

(11)特許出願公開番号

特開平11-212978

(43)公開日 平成11年(1999)8月6日

(51) Int.Cl.<sup>6</sup>

**識別記号**

FI

G O 6 F 17/30

**G O 6 F 15/401**

310D

審査請求 未請求 請求項の数23 O.L (全 20 頁)

(21)出願番号 特願平10-14314

(22)出願日 平成10年(1998)1月27日

(71)出願人 000006013

三菱電機株式会社

東京都千代田区丸の内二丁目2番3号

(72) 発明者 小中 裕喜

東京都千代田区丸の内二丁目2番3号 三  
菱電機株式会社内

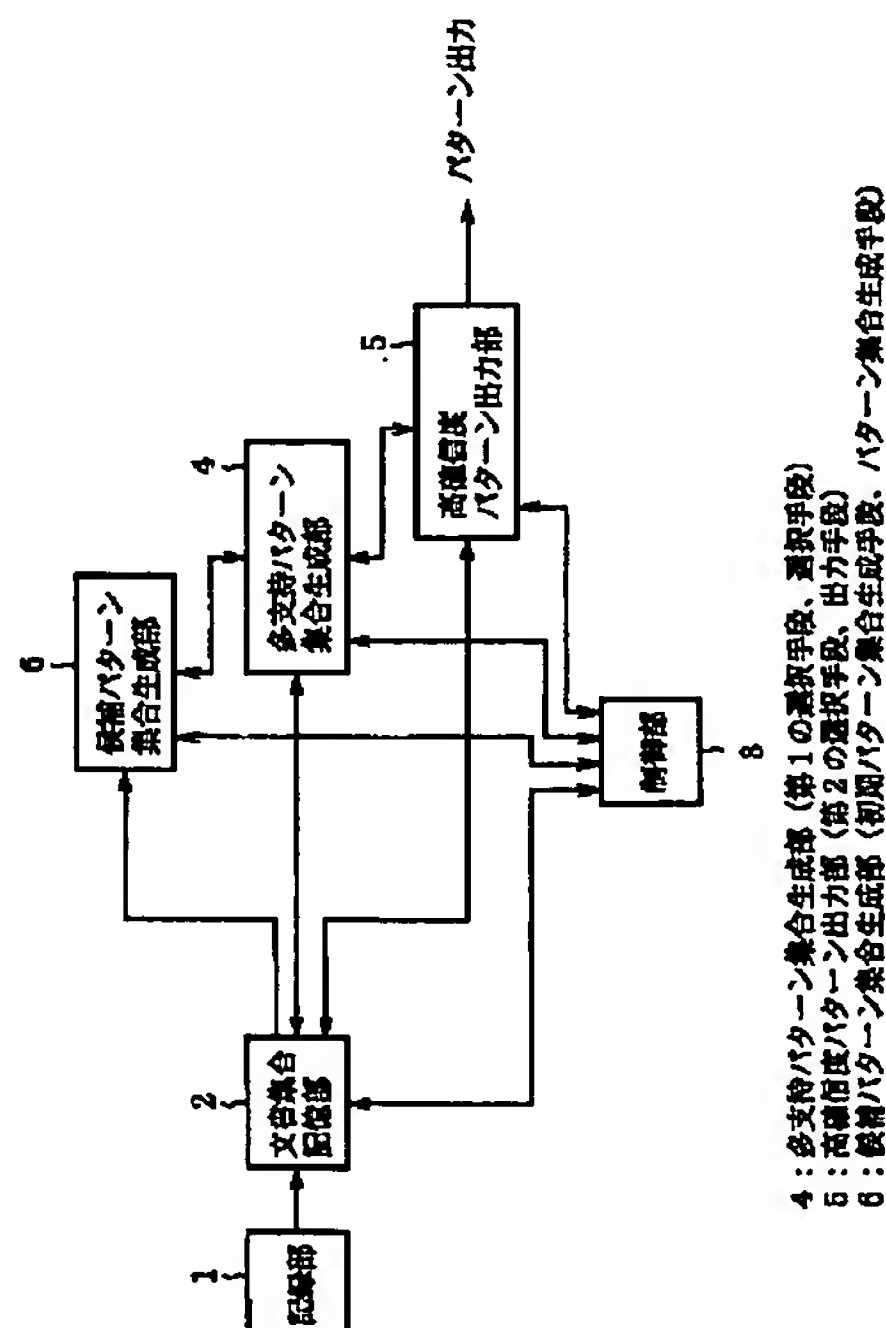
(74) 代理人 弁理士 田澤 博昭 (外1名)

(54) 【発明の名称】 文書情報解析方法、文書情報解析装置および記録媒体

(57) 【要約】

【課題】 他のクラスタと比較してクラスタ特有のパターンを抽出する。

【解決手段】 制御部3により指定されたクラスタに対応する各パターンについて、多支持パターン集合生成部4は、そのクラスタにおいてパターンが出現する文書情報の数とクラスタに属する全ての文書情報の数との比である支持率を計算し、その支持率が所定の最小支持率以上のパターンを選択し、高確信度パターン出力部5に出力する。高確信度パターン出力部5は、そのクラスタにおいてパターンが出現する文書情報の数と全てのクラスタにおいてパターンが出現する文書情報の数との比である確信度を計算し、確信度が所定の最小確信度以上であるパターンを選択する。



## 【特許請求の範囲】

【請求項 1】 複数のクラスタに分類された、文書 ID とその文書 ID に対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ 1 つのキーワードで構成されるパターンの集合を初期のパターン集合として生成するステップと、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択するステップと、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択するステップと、

前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力するステップと、

前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数を  $n$  としたとき、複数のパターンを構成するキーワードのうちの  $(n+1)$  個のキーワードで構成され、かつ、その  $(n+1)$  個のキーワードのうちの任意の  $n$  個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するステップとを備えた文書情報解析方法。

【請求項 2】 所定のパターンの確信度の計算において、前記所定のパターンが出現した文書情報の文書 ID を記憶し、

所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を  $(m+1)$  個とすると、 $m$  個のキーワードで構成されるパターンが出現した文書情報の文書 ID を読み出し、その文書情報においてだけ、前記  $(m+1)$  個のキーワードで構成される所定のパターンが出現する文書情報をカウントすることを特徴とする請求項 1 記載の文書情報解析方法。

【請求項 3】 所定のパターンの確信度の計算において、前記所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、

所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を  $(m+1)$  個とすると、 $m$  個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出

し、そのクラスタにおいてだけ、前記  $(m+1)$  個のキーワードで構成される所定のパターンが出現する文書情報をカウントすることを特徴とする請求項 1 記載の文書情報解析方法。

【請求項 4】 初期のパターン集合を生成する前に、少なくとも 1 つのクラスタにおいて支持率が所定の最小支持率以上となるキーワード以外のキーワードを、すべての文書情報から除去することを特徴とする請求項 1 から請求項 3 のうちのいずれか 1 項記載の文書情報解析方法。

【請求項 5】 各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を前記元の文書情報の代わりに使用することを特徴とする請求項 1 から請求項 4 のうちのいずれか 1 項記載の文書情報解析方法。

【請求項 6】 所定のパターン集合は、初期のパターン集合であることを特徴とする請求項 5 記載の文書情報解析方法。

【請求項 7】 新たなパターン集合を生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去することを特徴とする請求項 1 から請求項 6 のうちのいずれか 1 項記載の文書情報解析方法。

【請求項 8】 初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが前記高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する前記高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力するとともに、前記高確信度パターン集合に追加することを特徴とする請求項 1 から請求項 7 のうちのいずれか 1 項記載の文書情報解析方法。

【請求項 9】 多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、前記確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力することを特徴とする請求項 1 から請求項 7 のうちのいずれか 1 項記載の文書情報解析方法。

【請求項 10】 クラスタ毎に一括して、所定の順番でパターンを出力することを特徴とする請求項 8 または請求項 9 記載の文書情報解析方法。

【請求項 11】 複数のクラスタに分類された、文書 ID とその文書 ID に対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキ

10

20

30

40

50

ーワードを読み出し、そのキーワードのうちのそれぞれ 1つのキーワードで構成されるパターン集合を初期のパターン集合として生成するステップと、  
 所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加するステップと、  
 前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数を  $n$  としたとき、複数の多支持パターンを構成するキーワードのうちの  $(n+1)$  個のキーワードで構成され、かつ、その  $(n+1)$  個のキーワードのうちの任意の  $n$  個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するステップと、  
 前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するステップとを備えた文書情報解析方法。

【請求項 12】 初期の多支持パターン集合を生成する前に、少なくとも 1つのクラスタにおいて支持率が所定の最小支持率以上となるキーワード以外のキーワードを、すべての文書情報から除去することを特徴とする請求項 11 記載の文書情報解析方法。

【請求項 13】 累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を前記元の文書情報の代わりに使用することを特徴とする請求項 11 または請求項 12 記載の文書情報解析方法。

【請求項 14】 クラスタ毎に一括して、所定の順番で、確信度が所定の最小確信度以上である多支持パターンを出力することを特徴とする請求項 11 から請求項 13 のうちのいずれか 1 項記載の文書情報解析方法。

【請求項 15】 所定の順番は、支持率の値に応じた順番であることを特徴とする請求項 10 または請求項 14 記載の文書情報解析方法。

【請求項 16】 所定の順番は、確信度の値に応じた順

番であることを特徴とする請求項 10 または請求項 14 記載の文書情報解析方法。

【請求項 17】 所定の順番は、パターンを構成するキーワードの数に応じた順番であることを特徴とする請求項 10 または請求項 14 記載の文書情報解析方法。

【請求項 18】 所定の順番は、各パターンの条件付きエントロピーの値に応じた順番であることを特徴とする請求項 10 または請求項 14 記載の文書情報解析方法。

【請求項 19】 すべてのクラスタから、パターンを出力するクラスタを選択するステップを備え、  
 選択したクラスタだけについてパターンを選択し、出力することを特徴とする請求項 1 から請求項 18 のうちのいずれか 1 項記載の文書情報解析方法。

【請求項 20】 複数のクラスタに分類された、文書 ID とその文書 ID に対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ 1つのキーワードで構成されるパターン集合を初期のパターン集合として生成する初期パターン集合生成手段と、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第 1 の選択手段と、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第 2 の選択手段と、

前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力する出力手段と、

前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数を  $n$  としたとき、複数のパターンを構成するキーワードのうちの  $(n+1)$  個のキーワードで構成され、かつ、その  $(n+1)$  個のキーワードのうちの任意の  $n$  個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段とを備えた文書情報解析装置。

【請求項 21】 複数のクラスタに分類された、文書 ID とその文書 ID に対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ



1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段と、前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数を $n$ としたとき、複数の多支持パターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段と、

前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段とを備えた文書情報解析装置。

【請求項 2 2】 コンピュータを、複数のクラスタに分類された、文書 ID とその文書 ID に対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ 1 つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第 1 の選択手段、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、前記所定のクラスタに対応するパターン

集合から、その確信度が所定の最小確信度以上であるパターンを選択する第 2 の選択手段、

前記支持率が所定の最小支持率以上であり、かつ、前記確信度が所定の最小確信度以上であるパターンを前記所定のクラスタに関連づけて出力する出力手段、

前記支持率が前記最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数を $n$ としたとき、複数のパターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが前記多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段として機能させるためのプログラムを記録した記録媒体。

【請求項 2 3】 コンピュータを、

複数のクラスタに分類された、文書 ID とその文書 ID に対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ 1 つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、

所定のクラスタに対応するパターン集合を構成する各パターンについて、前記所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、前記所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段、

前記多支持パターン集合において、前記多支持パターンを構成するキーワードの数を $n$ としたとき、複数の多支持パターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが前記多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段、

前記多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、前記累積多支持パターン集合を構成する各多支持パターンについて、前記所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、前記累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段として機能させるためのプログラムを記録した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、複数のクラスタに分類して記録されている、文書IDとその文書IDに対応するキーワードとで構成される文書情報において、各クラスタについて、そのクラスタを特徴づけるキーワードの連言（パターン）を抽出するための文書情報解析方法、文書情報解析装置および記録媒体に関するものである。

【0002】

【従来の技術】近年の情報通信メディア、特にインターネットやWWW（World Wide Web）などの発展に伴い、容易に入手可能な情報の量が急激に増加している。例えば、WWWにおいて各種情報を検索するために、Yahoo（商標）などの検索エンジンが提供されている。

【0003】従来、文書データベースにおいて検索エンジンなどを利用して所望の文書情報を検索する場合、1つまたは複数のキーワードを指定し、そのキーワードに該当する文書情報を文書データベースから読み出すようにしている。

【0004】しかしながら、大量の種々雑多な文書情報がそのまま記録されている場合、その文書情報の中から所望の文書情報を抽出するために、すべての文書情報を検索する必要があるが実用的ではない。例えば、検索対象に対する利用者の知識が曖昧であり、利用者がその対象を的確にとらえるキーワードを想起することが困難である場合、利用者は抽象的なキーワードを使用して検索するため、関連性の低い長大な文書情報が抽出されることが多く、利用者にとって興味のある文書情報に到達するまでに長い時間を費やしてしまうことになる。

【0005】そこで、オペレータが予めそのような種々雑多な文書情報をその内容に基づいて分類し、内容に関連性があるもの同士をクラスタごとにまとめて記録することや、SOM（Self-Organizing Maps）などの自己組織化アルゴリズムを使用して、種々雑多な文書情報をクラスタに分類することが考えられる。なお、自己組織化アルゴリズムには、例えば「Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration」（Lagusら著、「Proc. of the Second International Conference on Knowledge Discovery and Data Mining KDD'96」の第238頁～第243頁、1996年）に記載のものや、「Exploration of Document Collections with Self-Organizing Maps: A Novel Approach to Similarity Representation」（Merkl著、「Proc. of the first European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'97」の第101頁～第111頁、1997年）に記載のものがある。

【0006】このように、文書情報をクラスタに分類しておくことにより、所望の文書情報を検索する場合、まず、その文書情報に関連するクラスタを検索し、該当す

るクラスタだけを検査すればよいことになる。

【0007】しかしながら、文書情報をその詳細な内容にわたってクラスタリングする場合、オペレータによるクラスタリングでは、時間やコストが多く必要になることになる。一方、SOMなどの自己組織化アルゴリズムを使用したクラスタリングでは、各クラスタがどのような特徴に基づいて生成されたのか、また各クラスタの文書情報がどのような理由でそのクラスタに分類されたのかを判断することが困難な場合が多い。

【0008】近年、データベースから特徴的なパターンを抽出するいわゆるデータマイニング技術が提案されている。このデータマイニング技術により、所定の文書情報から、その文書情報において特徴的なキーワードのパターンを抽出することができる。したがって、特徴的なキーワードのパターンを抽出しておき、その文書情報に関連づけて記録しておくことにより、そのパターンに基づいて文書情報の検索を行えば、比較的容易に所望の文書情報に到達することになると考えられる。

【0009】そのようなデータマイニングの方法としては、例えば、「Fast Algorithms for Mining Association Rules」（Agrawalら著、「Proc. of VLDB'94」の第487頁～第499頁、1994年）（以下、文献1という）に記載のものや「An Effective Hash Based Algorithm for Mining Association Rules」（Parkら著、「Proc. of SIGMOD'95」の第175頁～第186頁、1995年）（以下、文献2という）に記載のものがある。さらに文書情報に適用したものとしては、「Pattern Based Browsing in Document Collections」（Feldmanら著、「Proc. of the First European Symposium on Principles of Data Mining and Knowledge Discovery PKDD'97」の第112頁～第122頁、1997年）（以下、文献3という）に記載のものがある。

【0010】

【発明が解決しようとする課題】文書情報から特徴的なキーワードのパターンを抽出する従来の文書解析方法は以上のように構成されているので、例えば文献3に記載の方法により、個々のクラスタ内において特徴的なパターンを抽出することはできるが、他のクラスタのパターンと比較して特徴的なそのクラスタ特有のパターンを抽出することが困難であるなどの課題があった。

【0011】すなわち、複数の類似するクラスタが存在する場合に、文書情報の検索を容易にするようなそのクラスタ特有のパターンを抽出することが困難であるという課題があった。

【0012】例えば、図9は、スポーツ関係の新聞記事を記録したデータベースの一例を示している。図9に示すデータベースにおいては、サッカー関連の文書情報と、高校総体関連の文書情報が、各クラスタに対応するファイルに記録されている。この場合、例えば「つり



輪」という単一のキーワードにより構成されるパターンは、「高校総体総合」クラスと「高校総体つり輪」クラスにおいて多く出現するので、このパターンにより、「高校総体総合」クラスまたは「高校総体つり輪」クラスを特徴づけることは困難である。一方、例えば「つり輪」と「種目別」により構成されるパターンは、「高校総体つり輪」クラスにおいて多く出現するので、このパターンにより、「高校総体つり輪」クラスを特徴づけることが可能である。しかしながら、この

ような、クラス特有のパターン、例えば上述の「つり輪」と「種目別」とにより構成されるパターンを抽出することは、個々のクラス内だけにおいて特徴的なパターンを抽出する従来の方法では困難であった。

【0013】この発明は上記のような課題を解決するためになされたもので、複数のクラスに分類して記録されている、文書IDとその文書IDに対応するキーワードとで構成される文書情報において、各クラスについて、そのクラスを特徴づけるキーワードのパターンを抽出するための文書情報解析方法、文書情報解析装置および記録媒体を得ることを目的とする。

【0014】

【課題を解決するための手段】この発明に係る文書情報解析方法は、複数のクラスに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラス毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスに対応するパターン集合を構成する各パターンについて、所定のクラスにおいてそのパターンが出現する文書情報の数とそのクラスに属するすべての文書情報の数との比である支持率を計算し、所定のクラスに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、所定のクラスに対応するパターン集合を構成する各パターンについて、所定のクラスにおいてそのパターンが出現する文書情報の数とすべてのクラスにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択し、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスに関連づけて出力し、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数を $n$ としたとき、複数のパターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するものである。

【0015】この発明に係る文書情報解析方法は、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の文書IDを記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を $(m+1)$ 個とすると、 $m$ 個のキーワードで構成されるパターンが出現した文書情報の文書IDを読み出し、その文書情報においてだけ、 $(m+1)$ 個のキーワードで構成される所定のパターンが出現する文書情報をカウントするものである。

10 【0016】この発明に係る文書情報解析方法は、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の属するクラスの情報を読み出し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を $(m+1)$ 個とすると、 $m$ 個のキーワードで構成されるパターンが出現した文書情報の属するクラスの情報を読み出し、そのクラスにおいてだけ、 $(m+1)$ 個のキーワードで構成される所定のパターンが出現する文書情報をカウントするものである。

20 【0017】この発明に係る文書情報解析方法は、初期のパターン集合を生成する前に、少なくとも1つのクラスにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを、すべての文書情報から除去するものである。

【0018】この発明に係る文書情報解析方法は、各クラスについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

30 【0019】この発明に係る文書情報解析方法は、各クラスについて、初期のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

40 【0020】この発明に係る文書情報解析方法は、新たなパターン集合を生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去するものである。

50 【0021】この発明に係る文書情報解析方法は、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけ

を、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加するものである。

【0022】この発明に係る文書情報解析方法は、多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、その確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力するものである。

【0023】この発明に係る文書情報解析方法は、クラスタ毎に一括して、所定の順番でパターンを出力するものである。

【0024】この発明に係る文書情報解析方法は、複数のクラスタに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加し、多支持パターン集合において、多支持パターンを構成するキーワードの数を $n$ としたとき、複数の多支持パターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するものである。

【0025】この発明に係る文書情報解析方法は、初期のパターン集合を生成する前に、少なくとも1つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードをすべての文書情報から除去するものである。

【0026】この発明に係る文書情報解析方法は、累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラ

スタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するものである。

【0027】この発明に係る文書情報解析方法は、クラスタ毎に一括して、所定の順番で、確信度が所定の最小確信度以上であるパターンを出力するものである。

【0028】この発明に係る文書情報解析方法は、支持率の値に応じた順番でパターンを出力するものである。

【0029】この発明に係る文書情報解析方法は、確信度の値に応じた順番でパターンを出力するものである。

【0030】この発明に係る文書情報解析方法は、パターンを構成するキーワードの数に応じた順番でパターンを出力するものである。

【0031】この発明に係る文書情報解析方法は、各パターンの条件付きエントロピーの値に応じた順番でパターンを出力するものである。

【0032】この発明に係る文書情報解析方法は、すべてのクラスタから選択されたクラスタだけについてパターンを選択し、出力するものである。

【0033】この発明に係る文書情報解析装置は、複数のクラスタに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第1の選択手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第2の選択手段と、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力する出力手段と、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数を $n$ としたとき、複数のパターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段とを備えたものであ



る。

【0034】この発明に係る文書情報解析装置は、複数のクラスタに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段と、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段と、多支持パターン集合において、多支持パターンを構成するキーワードの数を $n$ としたとき、複数の多支持パターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段と、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段とを備えたものである。

【0035】この発明に係る記録媒体は、コンピュータを、複数のクラスタに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択する第1の選択手段、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパ

ターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択する第2の選択手段、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力する出力手段、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数を $n$ としたとき、複数のパターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するパターン集合生成手段として機能させるためのプログラムを記録したものである。

【0036】この発明に係る記録媒体は、コンピュータを、複数のクラスタに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成する初期パターン集合生成手段、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加する選択手段、多支持パターン集合において、多支持パターンを構成するキーワードの数を $n$ としたとき、複数の多支持パターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成するパターン集合生成手段、多支持パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力する出力手段として機能させるためのプログラムを記録したものである。

【0037】

【発明の実施の形態】以下、この発明の実施の一形態を



説明する。

実施の形態 1. 図 1 は、この発明の実施の形態 1 による文書情報解析装置の構成を示すブロック図である。図において、1 は、例えば図 9 に示すような文書情報をクラスタに分類して、各クラスタに対応するファイルにそれぞれ対応する文書情報を記録する例えばハードディスク装置、フロッピーディスクを装着したフロッピーディスク駆動装置などの記録部である。なお、文書情報は、図 9 に示すように文書 ID と、1 つまたは複数のキーワードで構成される。

【0038】2 は、記録部 1 に記録された文書情報をクラスタ毎に読み出して記憶するとともに、制御部 3 により指定されたクラスタに対応する文書情報を候補パターン生成部 6 に出力する文書集合記憶部である。また文書集合記憶部 2 は、多支持パターン集合生成部 4 または高確信度パターン出力部 5 によりクラスタとパターンが指定されると、指定されたクラスタにおいて、指定されたパターンが出現する文書情報の数を計算し、多支持パターン集合生成部 4 または高確信度パターン出力部 5 に出力する。なお、文書集合記憶部 2 は、各文書情報において、指定されたパターンが出現するかどうかを検査する場合、その文書情報を構成するキーワードの部分集合がそのパターンに一致するかどうかを検査する。したがって、この検査を効率的に行うために、ハッシュテーブルやハッシュ木を用いて各文書のキーワード集合に対して部分集合となる可能性のあるパターンを絞り込んだり、各パターン、文書情報においてキーワード集合を整理したり、あるいはビットパターンで表すようにしてもよい。なお、文書集合記憶部 2 を所定のコンピュータネットワークに接続し、そのコンピュータネットワークを介して他の記録部から文書情報を読み出すようにしてもよい。

【0039】6 は、文書集合記憶部 2 より供給されたあるクラスタの文書情報に用いられている各キーワードに対し、1 つのキーワードにより構成される候補パターンを生成して、多支持パターン集合生成部 4 に出力するとともに、多支持パターン集合生成部 4 より供給され、その多支持パターン集合の多支持パターンを構成するキーワードの数を  $n$  としたとき、 $(n-1)$  個のキーワードが共通する任意の 2 つの多支持パターンから  $(n+1)$  個のキーワードからなるパターンを生成し、そのパターンの任意の  $n$  個のキーワードが多支持パターン集合に含まれるものを候補パターンとして、新たな候補パターン集合を生成し、多支持パターン集合生成部 4 に出力する候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）である。

【0040】4 は、候補パターン集合生成部 6 より供給される各候補パターンについて、制御部 3 により指定されたクラスタにおいてその候補パターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を、文書集合記憶部 2 を利用して計

算し、その支持率が所定の最小支持率以上であるパターンを新たな多支持パターンとして選択し、それらの多支持パターン集合を高確信度パターン出力部 5 と候補パターン集合生成部 6 に出力する多支持パターン集合生成部（第 1 の選択手段、選択手段）である。

【0041】5 は、多支持パターン集合生成部 4 より供給される各パターンについて、制御部 3 により指定されたクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を、文書集合記憶部 2 を利用して計算し、多支持パターン集合生成部 4 より供給されたパターン集合から、確信度が所定の最小確信度以上であるパターンを選択し、そのパターンを図示せぬ装置に出力する高確信度パターン出力部（第 2 の選択手段、出力手段）である。なお、高確信度パターン出力部 5 は、図示せぬ装置に出力するかわりに、そのパターンを、指定されたクラスタに関連づけて記録部 1 に記録させるようにしてもよい。

【0042】3 は、文書集合記憶部 2、多支持パターン集合生成部 4、高確信度パターン出力部 5、および候補パターン集合生成部 6 を制御する制御部である。

【0043】なお、上述の文書集合記憶部 2、制御部 3、多支持パターン集合生成部 4、高確信度パターン出力部 5、および候補パターン集合生成部 6 は、各種電子回路でハードウェアとして構成してもよいが、CPU を文書集合記憶部 2、制御部 3、多支持パターン集合生成部 4、高確信度パターン出力部 5、および候補パターン集合生成部 6 として機能させるためのプログラムを記録した記録媒体（例えば、ハードディスク装置、フロッピーディスクなどの磁気ディスク、コンパクトディスクなどの光ディスク）から、そのプログラムを読み取ることが可能なコンピュータで構成してもよい。

【0044】次に動作について説明する。図 2 は、図 1 の文書情報解析装置の動作を説明するフローチャートである。まず、ステップ ST1 において、文書集合記憶部 2 は、記録部 1 に記録された文書情報を読み出し、候補パターン集合生成部 6 に出力する。

【0045】ステップ ST2 において、制御部 3 は、文書集合記憶部 2 により文書情報を読み出されたクラスタのうち、後述のステップ ST3～ステップ ST10 の処理を実行されていないクラスタがあるかどうかを判定し、すべてのクラスタについてその処理を実行したと判定した場合、パターン抽出の処理を終了し、その処理を実行されていないクラスタが少なくとも 1 つあると判定した場合、そのクラスタのうちの 1 つを指定する。

【0046】候補パターン集合生成部 6 は、ステップ ST3 において、供給された文書情報に含まれる各キーワードを 1 つのキーワードからなる候補パターン（初期候補パターン）とし、クラスタ毎に候補パターン集合（初期候補パターン集合）を生成し、多支持パターン集合生

10

20

30

40

50

成部4に出力する。そして、ステップST4において、多支持パターン集合生成部4は、制御部3により指定されたクラスタに属する各候補パターンについて、文書集合記憶部2を制御してそのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップST5において、そのクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率をそれぞれ計算し、その支持率が所定の最小支持率以上であるパターンを多支持パターンとして選択し、それらの多支持パターンの集合を多支持パターン集合として高確信度パターン出力部5と候補パターン生成部6に出力する。

【0047】次に、高確信度パターン出力部5は、ステップST6において、多支持パターン集合生成部4より供給された多支持パターンについて、文書集合記憶部2を制御して、指定されたクラスタ以外の残りのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップST7において、そのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、その確信度が所定の最小確信度以上であるパターンを選択し、図示せぬ装置に出力する。すなわち、指定されたクラスタにおいてそのパターンが出現する文書情報の数と、すべてのクラスタにおいてそのパターンが出現する文書情報の数とに基づいた確信度を利用することにより、そのクラスタ特有のパターンが選択される。

【0048】そして、ステップST8において、制御部3は、多支持パターン集合生成部4により選択された多支持パターンの数が所定の最小パターン数より小さいか否かを判定し、選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、指定したクラスタに対する処理を終了し、ステップST2に戻る。一方、選択されたパターンの数が所定の最小パターン数以上であると判定した場合、制御部3は、ステップST9において、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したか否かを判定し、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、指定したクラスタに対する処理を終了し、ステップST2に戻る。

【0049】一方、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達していないと判定した場合、制御部3は、候補パターン集合生成部6に制御信号を供給する。ステップST10において、候補パターン集合生成部6は、その制御信号を受信すると、多支持パターン集合生成部4より供給された多支持パターン集合について、多支持パターンを構成するキーワードの数を $n$ としたとき、複数の多支持パターン

( $(n-1)$ 個のキーワードが共通である2つの多支持パターン)を構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキ

ーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、そのパターン集合を候補パターン集合として多支持パターン集合生成部4に出力する。なお、このステップST10の処理には、文献1に記載されている「apriori-gen」関数を利用することができる。

【0050】そして、ステップST4に戻り、ステップST10で生成された候補パターン集合について、ステップST4～ステップST10の処理が実行される。

【0051】すなわち、ステップST5～ステップST10の処理を反復する毎に、パターンを構成するキーワードの数を1つずつ増えていくとともに、それに伴い、支持率が所定の最小支持率以上であるパターンが次第に減少していく。

【0052】以上のように、この実施の形態1によれば、各クラスタについて、所定のパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率と、そのクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度とに基づいて、キーワードのパターンを抽出するので、そのクラスタ特有のパターンを抽出することができるという効果が得られる。

【0053】なお、ステップST4～ステップST10の反復処理において、2回目以降の処理では、前回の処理において選択された多支持パターンを有していた文書情報またはクラスタだけについて、候補パターンが出現するか否かの検査を行うようにしてもよい。前回の処理において選択された多支持パターンを有しない文書情報は、そのパターンを含む今回の処理の候補パターンも同様に有しないので、このようにすることにより、処理する文書情報やクラスタの数が低減され、処理時間を短縮することができるという効果が得られる。

【0054】なお、オペレータが図示せぬ操作部を操作して選択したクラスタだけについて、パターンの抽出の処理を行うようにしてもよい。その場合、図示せぬ操作部は、制御部3に接続され、その操作に対応する信号に応じて制御部3は、各部を制御する。このようにすることにより、不要なクラスタに対して処理が行われないので、処理時間を短縮することができるという効果が得られる。

【0055】実施の形態2. この発明の実施の形態2による文書情報解析装置は、実施の形態1による文書情報解析装置の文書集合記憶部2と多支持パターン集合生成部4に後述の機能を追加したものである。

【0056】すなわち、実施の形態2における文書集合記憶部2は、記録部1から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情

10

20

30

40

50



報から除去する。また、実施の形態2における多支持パターン集合生成部4は、候補パターン集合生成部6より供給される候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【0057】なお、その他の構成要素は、実施の形態1のものと同様であるので、その説明を省略する。

【0058】次に動作について説明する。図3は、実施の形態2による文書情報解析装置の動作を説明するフローチャートである。なお、ステップST1～ステップST10の処理は、実施の形態1のものと同様であるので、その説明を省略する。

【0059】実施の形態2においては、ステップST1の後のステップST21において、文書集合記憶部2は、記録部1から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。

【0060】また、ステップST5の後のステップST22において、多支持パターン集合生成部4は、候補パターン集合生成部6より供給された候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【0061】なお、ステップST22において更新された文書情報は、各クラスタについての処理が終了したときに消去され、次のクラスタについての処理は、ステップST21において生成された文書情報に基づいて行われる。

【0062】以上のように、この実施の形態2によれば、パターンを構成する各キーワードについて支持率を計算し、その支持率が所定の最小支持率より低いキーワードを文書情報から適宜除去するようにしたので、パターンが出現するか否かをチェックする対象になる文書情報の量を低減させることができ、処理時間を短縮することができるという効果が得られる。

【0063】なお、ステップST22の処理は、ステップST22、ステップST4～ステップST10の反復処理の最初の処理においてだけ行うようにしてもよい。2回目以降のステップST22の処理においては、除去されるキーワードの数が少ないので、このようにすることにより、より処理時間を短縮することができる場合が多い。

【0064】実施の形態3. この発明の実施の形態3による文書情報解析装置は、実施の形態2による文書情報解析装置の高確信度パターン出力部5に後述の機能を追加したものである。

【0065】すなわち、実施の形態3における高確信度パターン出力部5は、多支持パターン集合生成部4より供給されたパターンから、その確信度が所定の最小確信度以上であるパターンを選択し、そのパターンを図示せぬ装置に出力した後、確信度が所定の最大確信度以上であるパターンを除去し、残りのパターンを新たに多支持パターン集合として多支持パターン集合生成部4に戻す。なお、多支持パターン集合生成部4は、高確信度パターン出力部5から戻された多支持パターン集合を候補パターン集合生成部6に出力する。

【0066】なお、その他の構成要素は、実施の形態2のものと同様であるので、その説明を省略する。

【0067】次に動作について説明する。図4は、実施の形態3による文書情報解析装置の動作を説明するフローチャートである。なお、ステップST1～ステップST10、ステップST21、およびステップST22の処理は、実施の形態2のものと同様であるので、その説明を省略する。

【0068】実施の形態3においては、高確信度パターン出力部5は、ステップST7において、多支持パターン集合生成部4より供給されたパターンから、その確信度が所定の最小確信度以上であるパターンを選択し、図示せぬ装置に出力した後、ステップST31において、その確信度が所定の最大確信度以上であるパターンを除去し、残りのパターンを新たに多支持パターン集合として多支持パターン集合生成部4に戻す。なお、多支持パターン集合生成部4は、ステップST6で多支持パターン集合を候補パターン集合生成部6に出力する代わりに、ステップST31で高確信度パターン出力部5より戻された多支持パターン集合を出力する。

【0069】以上のように、この実施の形態3によれば、確信度が所定の最大確信度以上であるパターンを多支持パターン集合から除去するようにしたので、そのようなパターンを含む、利用者にとって不必要に複雑なパターンの出力と、そのための処理を抑制することができるという効果が得られる。

【0070】実施の形態4. この発明の実施の形態4による文書情報解析装置は、実施の形態3による文書情報解析装置の高確信度パターン出力部5に後述の機能を追加したものである。

【0071】実施の形態4においては、高確信度パターン出力部5は、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、確信度がそのパターンのサブセットに該当する高確信度パターン集合におけるパターンの確信度以上であるパターンだけを、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加する。

【0072】なお、その他の構成要素は、実施の形態3のものと同様であるので、その説明を省略する。

【0073】次に動作について説明する。図5は、実施の形態4による文書情報解析装置の動作を説明するフローチャートである。なお、ステップST1～ステップST10、ステップST21、ステップST22、およびステップST31の処理は、実施の形態3のものと同様であるので、その説明を省略する。

【0074】実施の形態4においては、高確信度パターン出力部5は、ステップST41において、初期状態の空集合である高確信度パターン集合を記憶しておき、ステップST42において、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力し、その後、高確信度パターン集合に追加する。

【0075】以上のように、この実施の形態4によれば、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを選択するようにしたので、キーワード数が多くかつ確信度の低いパターンを除去することができ、キーワード数が少なく確信度の高いパターンが選択され、検索時の効率を向上させることができるという効果が得られる。

【0076】なお、ステップST10において新たなパターン集合を生成するときに、各パターンのすべてのサブセットの確信度を同時に計算し、そのパターンのすべてのサブセットの確信度のうちの最大値をそのパターンに対応して記憶して、ステップST42においては、そのサブセットの確信度の最大値以上の確信度を有するパターンだけを選択し、そのパターンを出力するとともに、高確信度パターン集合に追加するようにしてもよい。このようにすることにより、ステップST42における処理を簡略化することができ、処理時間を短縮することができるという効果が得られる。

【0077】実施の形態5、この発明の実施の形態5による文書情報解析装置は、実施の形態4による文書情報解析装置の高確信度パターン出力部5に後述の機能を追加したものである。

【0078】すなわち、実施の形態5においては、高確信度パターン出力部5は、支持率が所定の最小支持率以

上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、そのパターンの確信度がそのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上であるパターンだけ高確信度パターン集合に追加していき、各クラスタに対する処理が終了したとき、クラスタ毎に一括して、所定の順番で高確信度パターン集合を出力する。

10 【0079】なお、その他の構成要素は、実施の形態4のものと同様であるので、その説明を省略する。

【0080】次に動作について説明する。図6は、実施の形態5による文書情報解析装置の動作を説明するフローチャートである。なお、ステップST1～ステップST10、ステップST21、ステップST22、ステップST31、ステップST41、およびステップST42の処理は、実施の形態4のものと同様であるので、その説明を省略する。

20 【0081】実施の形態5においては、高確信度パターン出力部5は、ステップST42において、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないか、あるいは、そのパターンの確信度がそのパターンのサブセットに該当する高確信度パターン集合のパターンの確信度以上であるパターンだけ高確信度パターン集合に追加していく。なお、このときには、高確信度パターン出力部5は、そのパターンを図示せぬ所定の装置には出力しない。

30 【0082】そして、高確信度パターン出力部5は、ステップST8またはステップST9において、選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、または、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、ステップST51において、そのクラスタに対応するパターンとして、所定の順番で高確信度パターン集合を出力する。

40 【0083】なお、このときの順番は、高確信度パターン集合を構成するパターンの支持率、確信度、パターンを構成するキーワードの数などに基づいて予め設定する。

【0084】また、Dを文書情報の集合、D<sub>p</sub>をパターンpが出現する文書情報の集合としたときの、式(1)および式(2)により規定される各パターンpに対する所定のクラスタc<sub>i</sub>の条件付きエントロピーEnt<sub>i</sub>(D|p)の値に基づいて順番を設定してもよい。

【数1】



23

$$\begin{aligned} \text{Ent}_i(D|P) = & P(D_p) \times \text{Ent}_i(D_p) \\ & + P(\overline{D_p}) \times \text{Ent}_i(\overline{D_p}) \quad \dots (1) \end{aligned}$$

$$\text{ただし、} \overline{D_p} = D - D_p, P(D_p) = |D_p| / |D|$$

【0085】

\* \* 【数2】

$$\text{Ent}_i(D) = - (P_i^+ \log P_i^+ + P_i^- \log P_i^-) \quad \dots (2)$$

$$\text{ただし、} P_i^+ = |C_i \cap D| / |D|, P_i^- = 1 - P_i^+$$

【0086】以上のように、この実施の形態5によれば、所定の順番でパターンを出力するようにしたので、利用者の要求に応じた順番でパターンを出力することができるという効果が得られる。

【0087】実施の形態6. この発明の実施の形態6による文書情報解析装置は、実施の形態1による文書情報解析装置の制御部3による制御手順を後述のように変更したものである。

【0088】なお、その他の構成要素は、実施の形態1のものと同様であるので、その説明を省略する。

【0089】次に動作について説明する。図7は、実施の形態6による文書情報解析装置の動作を説明するフローチャートである。まず、ステップST1において、文書集合記憶部2は、記録部1に記録された文書情報を読み出し、候補パターン集合生成部6に出力する。

【0090】ステップST2において、制御部3は、文書集合記憶部2により文書情報を読み出されたクラスタのうち、後述のステップST3～ステップST5、ステップST8～ステップST10、ステップST61およびステップST63～ステップST65の処理を実行されていないクラスタがあるか否かを判定し、すべてのクラスタについてその処理を実行したと判定した場合、パターン抽出の処理を終了し、その処理を実行されていないクラスタが少なくとも1つあると判定した場合、そのクラスタのうちの1つを指定する。

【0091】候補パターン集合生成部6は、ステップST3において、供給された文書情報に含まれる各キーワードを1つのキーワードからなる候補パターン（初期候補パターン）とし、クラスタ毎に候補パターン集合（初期候補パターン集合）を生成し、多支持パターン集合生成部4に出力する。そして、ステップST4において、多支持パターン集合生成部4は、制御部3により指定されたクラスタに属する各候補パターンについて、文書集合記憶部2を制御してそのクラスタにおいてそのパターンが出現する文書情報の数を計算させ、ステップST5において、そのクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率をそれぞれ計算し、その支持率が所定の最小支持率以上であるパターンを多支持パターンとして選択し、それらの多支持パターンの集合を多

24

支持パターン集合として候補パターン生成部6に出力する。

【0092】そして、ステップST61において、多支持パターン集合生成部4は、多支持パターンを累積多支持パターン集合に追加する。なお、累積多支持パターン集合は、各クラスタについての処理を開始するとき（ステップST3）において、空集合に設定される。

【0093】そして、ステップST8において、制御部3は、多支持パターン集合生成部4により選択されたパターンの数が所定の最小パターン数より小さいか否かを判定し、選択されたパターンの数が所定の最小パターン数より小さくないと判定した場合、ステップST9に進み、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したか否かを判定する。

【0094】そして、ステップST8において選択されたパターンの数が所定の最小パターン数より小さいと判定した場合、または、ステップST9において選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達したと判定した場合、ステップST63に進み、制御部3は、多支持パターン集合生成部4に制御信号を供給し、そのときに選択されているパターンを高確信度パターン出力部5に出力させる。

【0095】一方、ステップST9において、選択されたパターンを構成するキーワードの数が所定の最大キーワード数に達していないと判定した場合、制御部3は特に何もしない。したがって、ステップST10に進み、候補パターン集合生成部6は、多支持パターン集合生成部4より供給された多支持パターン集合について、多支持パターンを構成するキーワードの数をnとしたとき、複数の多支持パターンを構成するキーワードのうちの(n+1)個のキーワードで構成され、その(n+1)個のキーワードのうちの任意のn個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターンの集合を生成し、そのパターン集合を候補パターン集合として多支持パターン集合生成部4に出力する。そして、ステップST4に戻り、ステップST10で生成された候補パターン集合について上述の処理を同様に実行する。

【0096】このようにして、ステップST8またはステップST9の条件が成立するまで、ステップST4、

ステップST5、ステップST61、ステップST8～ステップST10の処理が反復的に実行される。

【0097】そして、反復処理が終了した後、ステップST63においては、制御部3は、多支持パターン集合生成部4に制御信号を供給し、そのときに選択されているパターンを高確信度パターン出力部5に出力させる。

【0098】高確信度パターン出力部5は、ステップST64において、多支持パターン集合生成部4より供給された各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、多支持パターン集合生成部4より供給されたパターン集合から、確信度が所定の最小確信度以上であるパターンを選択し、ステップST65において、そのパターンを図示せぬ装置に出力する。そして、ステップST2に戻り、制御部3は、次のクラスタに対して同様の処理を行うように各部を制御する。

【0099】以上のように、この実施の形態6によれば、指定されたクラスタ以外のクラスタにおいてパターンが出現する文書情報を1回だけカウントすればよいので、処理時間を短縮することができるという効果が得られる。

【0100】なお、この実施の形態6においても、実施の形態5のように、所定の順番でパターンを出力するようにしてもよい。

【0101】実施の形態7. この発明の実施の形態7による文書情報解析装置は、実施の形態6による文書情報解析装置の文書集合記憶部2と多支持パターン集合生成部4に後述の機能を追加したものである。

【0102】すなわち、実施の形態7における文書集合記憶部2は、記録部1から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。また、多支持パターン集合生成部4は、候補パターン集合生成部6より供給される候補パターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用する。

【0103】なお、その他の構成要素は、実施の形態6のものと同様であるので、その説明を省略する。

【0104】次に動作について説明する。図8は、実施の形態7による文書情報解析装置の動作を説明するフローチャートである。なお、ステップST1～ステップST5、ステップST8～ステップST10、ステップST61、およびステップST63～ステップST65の処理は、実施の形態6のものと同様であるので、その説明を省略する。

【0105】実施の形態7においては、ステップST1

の後のステップST71において、文書集合記憶部2が、記録部1から読み出したキーワードの各クラスタにおける支持率を計算し、その支持率の最大値が所定の最小支持率より低いキーワードをすべての文書情報から除去する。

【0106】また、ステップST63の前のステップST72において、高確信度パターン出力部5が、多支持パターン集合生成部4より供給されるパターンの集合から支持率が所定の最小支持率以上であるパターンを選択した後、その選択したパターンの集合に含まれるキーワード以外のキーワードを除去した文書情報を生成する。そして、ステップST63の処理では、その文書情報が元の文書情報の代わりに使用される。

【0107】以上のように、この実施の形態7によれば、パターンを構成する各キーワードについて支持率を計算し、その支持率が所定の最小支持率より低いキーワードをすべての文書情報から除去するようにしたので、パターンが出現するか否かを検査する対象になる文書情報の数を低減させることができ、処理時間を短縮することができるという効果が得られる。

【0108】

【発明の効果】以上のように、この発明によれば、複数のクラスタに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とすべてのクラスタにおいてそのパターンが出現する文書情報の数との比である確信度を計算し、所定のクラスタに対応するパターン集合から、その確信度が所定の最小確信度以上であるパターンを選択し、支持率が所定の最小支持率以上であり、かつ、確信度が所定の最小確信度以上であるパターンを所定のクラスタに関連づけて出力し、支持率が所定の最小支持率以上であるパターンで構成される多支持パターン集合において、パターンを構成するキーワードの数を $n$ としたとき、複数のパターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかのパターンに該当する新たなパターン集合を生成するように構成したので、そのクラスタ特有のパターンを抽出することができ



る効果がある。

【0109】この発明によれば、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の文書IDを記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を $(m+1)$ 個とすると、 $m$ 個のキーワードで構成されるパターンが出現した文書情報の文書IDを読み出し、その文書情報においてだけ、 $(m+1)$ 個のキーワードで構成される所定のパターンが出現する文書情報をカウントするように構成したので、パターンの出現の検査の対象になる文書情報の数が低減され、処理時間を短縮することができる効果がある。

【0110】この発明によれば、所定のパターンの確信度の計算において、所定のパターンが出現した文書情報の属するクラスタの情報を記憶し、所定のパターンが出現する文書情報の数をカウントするとき、そのパターンを構成するキーワードの数を $(m+1)$ 個とすると、 $m$ 個のキーワードで構成されるパターンが出現した文書情報の属するクラスタの情報を読み出し、そのクラスタにおいてだけ、 $(m+1)$ 個のキーワードで構成される所定のパターンが出現する文書情報をカウントするように構成したので、パターンの出現の検査の対象になるクラスタの数が低減され、処理時間を短縮することができる効果がある。

【0111】この発明によれば、初期のパターン集合を生成する前に、少なくとも1つのクラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを、すべての文書情報から除去するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【0112】この発明によれば、各クラスタについて、所定のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【0113】この発明によれば、各クラスタについて、初期のパターン集合に属するすべてのパターンから支持率が所定の最小支持率以上であるパターンを選択した後に、元の文書情報から、そのパターン集合に属するパターンを構成するすべてのキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、処理の最初だけにおいて不要なキーワードが除去され、除去の処理にかかる時間を低減することができる効果がある。

【0114】この発明によれば、新たなパターン集合を

生成する前に、確信度が所定の最大確信度以上である多支持パターンを多支持パターン集合から除去するように構成したので、利用者の必要頻度が低く、かつそのクラスタにしか存在しないような複雑なパターンを除去することができる効果がある。

【0115】この発明によれば、初期状態では空集合である高確信度パターン集合を記憶し、支持率が所定の最小支持率以上であり、かつ確信度が所定の最小確信度以上であるパターンのうち、そのパターンのサブセットが高確信度パターン集合のいずれのパターンにも該当しないパターンと、そのパターンのサブセットのいずれかに該当する高確信度パターン集合のパターンの確信度以上である確信度を有するパターンとだけを、そのクラスタに関連づけて出力するとともに、高確信度パターン集合に追加するように構成したので、キーワード数が多くかつ確信度の低いパターンを除去することができ、キーワード数が少なく確信度の高いパターンが選択され、検索時の効率を向上させることができる効果がある。

【0116】この発明によれば、多支持パターン集合の各パターンについて、その確信度とともにそのパターンのすべてのサブセットの確信度を計算し、その確信度が所定の最小確信度以上でかつサブセットの確信度の最大値より大きい前記パターンだけを出力するように構成したので、キーワード数が長くかつ確信度の低いパターンを選択するときの処理を簡略化することができ、処理時間を短縮することができる効果がある。

【0117】この発明によれば、クラスタ毎に一括して、所定の順番でパターンを出力するように構成したので、利用者の要求に応じた順番でパターンを出力することができる効果がある。

【0118】この発明によれば、複数のクラスタに分類された、文書IDとその文書IDに対応するキーワードとで構成される文書情報を記録する記録部からクラスタ毎にすべてのキーワードを読み出し、そのキーワードのうちのそれぞれ1つのキーワードで構成されるパターンの集合を初期のパターン集合として生成し、所定のクラスタに対応するパターン集合を構成する各パターンについて、所定のクラスタにおいてそのパターンが出現する文書情報の数とそのクラスタに属するすべての文書情報の数との比である支持率を計算し、所定のクラスタに対応するパターン集合から、その支持率が所定の最小支持率以上であるパターンを選択し、そのパターンを多支持パターンとして多支持パターン集合および累積多支持パターン集合に追加し、多支持パターン集合において、多支持パターンを構成するキーワードの数を $n$ としたとき、複数の多支持パターンを構成するキーワードのうちの $(n+1)$ 個のキーワードで構成され、かつ、その $(n+1)$ 個のキーワードのうちの任意の $n$ 個のキーワードが多支持パターン集合のうちのいずれかの多支持パターンに該当する新たなパターン集合を生成し、多支持

パターン集合を構成する多支持パターンの数が所定の最小パターン数より小さいか、あるいは、多支持パターンを構成するキーワードの数が所定の最大キーワード数に達した場合、累積多支持パターン集合を構成する各多支持パターンについて、所定のクラスタにおいてその多支持パターンが出現する文書情報の数とすべてのクラスタにおいてその多支持パターンが出現する文書情報の数との比である確信度を計算し、累積多支持パターン集合に属する多支持パターンから、その確信度が所定の最小確信度以上である多支持パターンを選択し、出力するように構成したので、指定されたクラスタ以外のクラスタにおいてパターンが出現する文書情報を1回だけカウントすればよいので、処理時間を短縮することができる効果がある。

【0119】この発明によれば、累積多支持パターン集合を構成する各多支持パターンについて確信度を計算する前に、元の文書情報から、当該クラスタにおいて支持率が所定の最小支持率以上であるキーワード以外のキーワードを除去した文書情報を生成し、その文書情報を元の文書情報の代わりに使用するように構成したので、パターンの出現の検査の対象になるキーワードの数が低減され、処理時間を短縮することができる効果がある。

【0120】この発明によれば、すべてのクラスタから選択されたクラスタだけについてパターンを選択し、出力するように構成したので、不要なクラスタに対して処理が行われないので、処理時間を短縮することができる\*

\*効果がある。

【図面の簡単な説明】

【図1】 この発明の実施の形態1による文書情報解析装置の構成を示すブロック図である。

【図2】 図1の文書情報解析装置の動作を説明するフローチャートである。

【図3】 実施の形態2による文書情報解析装置の動作を説明するフローチャートである。

【図4】 実施の形態3による文書情報解析装置の動作を説明するフローチャートである。

【図5】 実施の形態4による文書情報解析装置の動作を説明するフローチャートである。

【図6】 実施の形態5による文書情報解析装置の動作を説明するフローチャートである。

【図7】 実施の形態6による文書情報解析装置の動作を説明するフローチャートである。

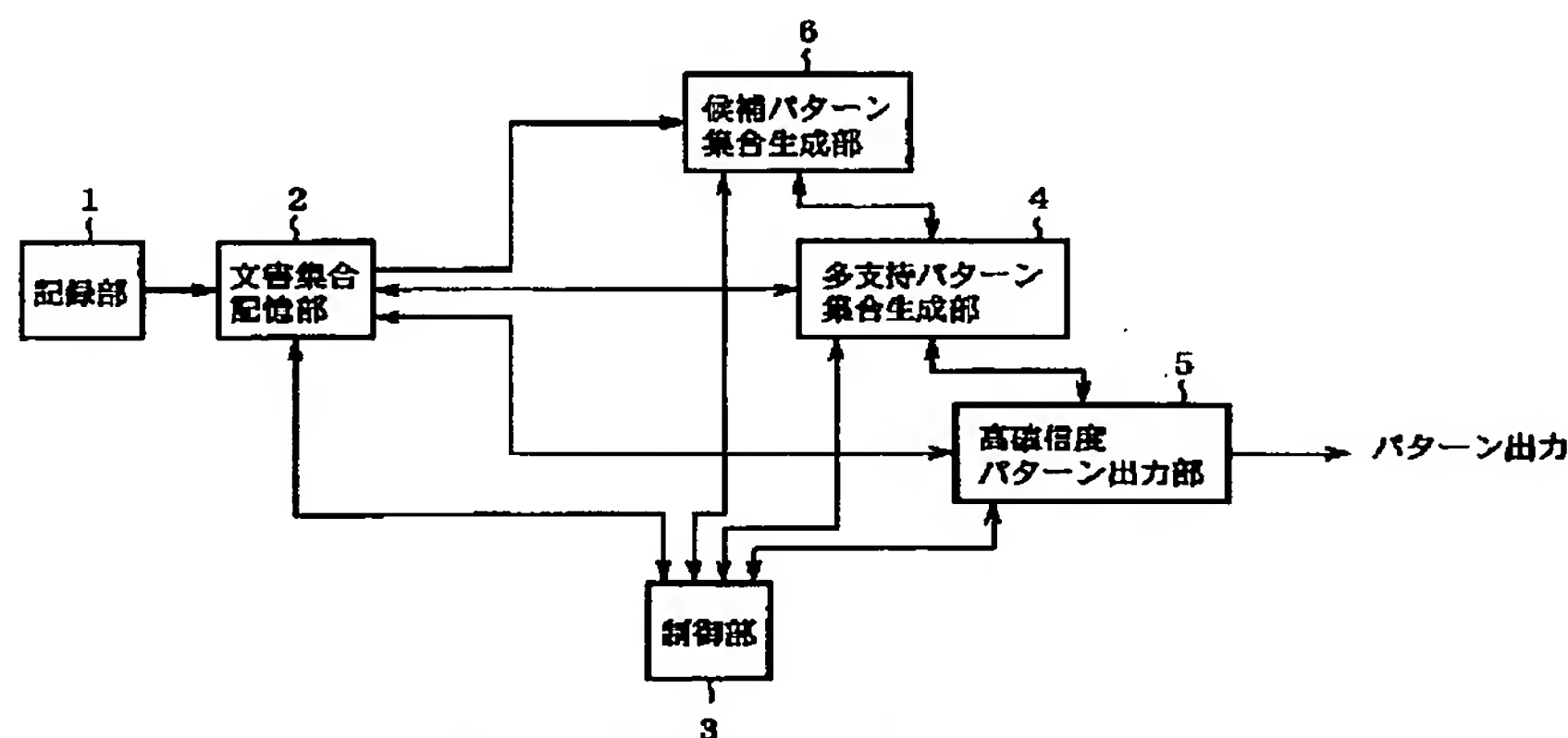
【図8】 実施の形態7による文書情報解析装置の動作を説明するフローチャートである。

【図9】 スポーツ関係の新聞記事を記録したデータベースの一例を示している。

【符号の説明】

1 記録部、4 多支持パターン集合生成部（第1の選択手段、選択手段）、5 高確信度パターン出力部（第2の選択手段、出力手段）、6 候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）。

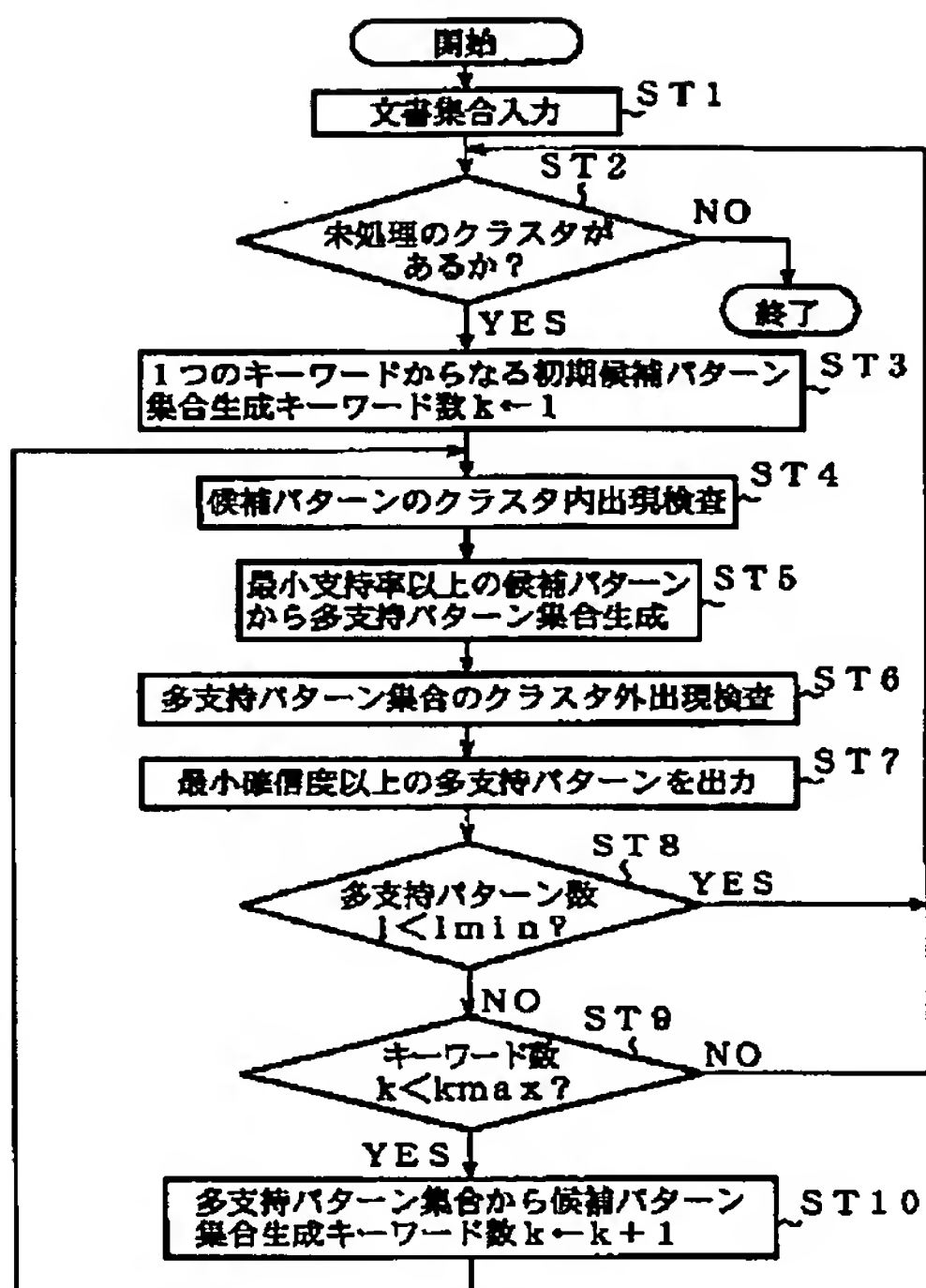
【図1】



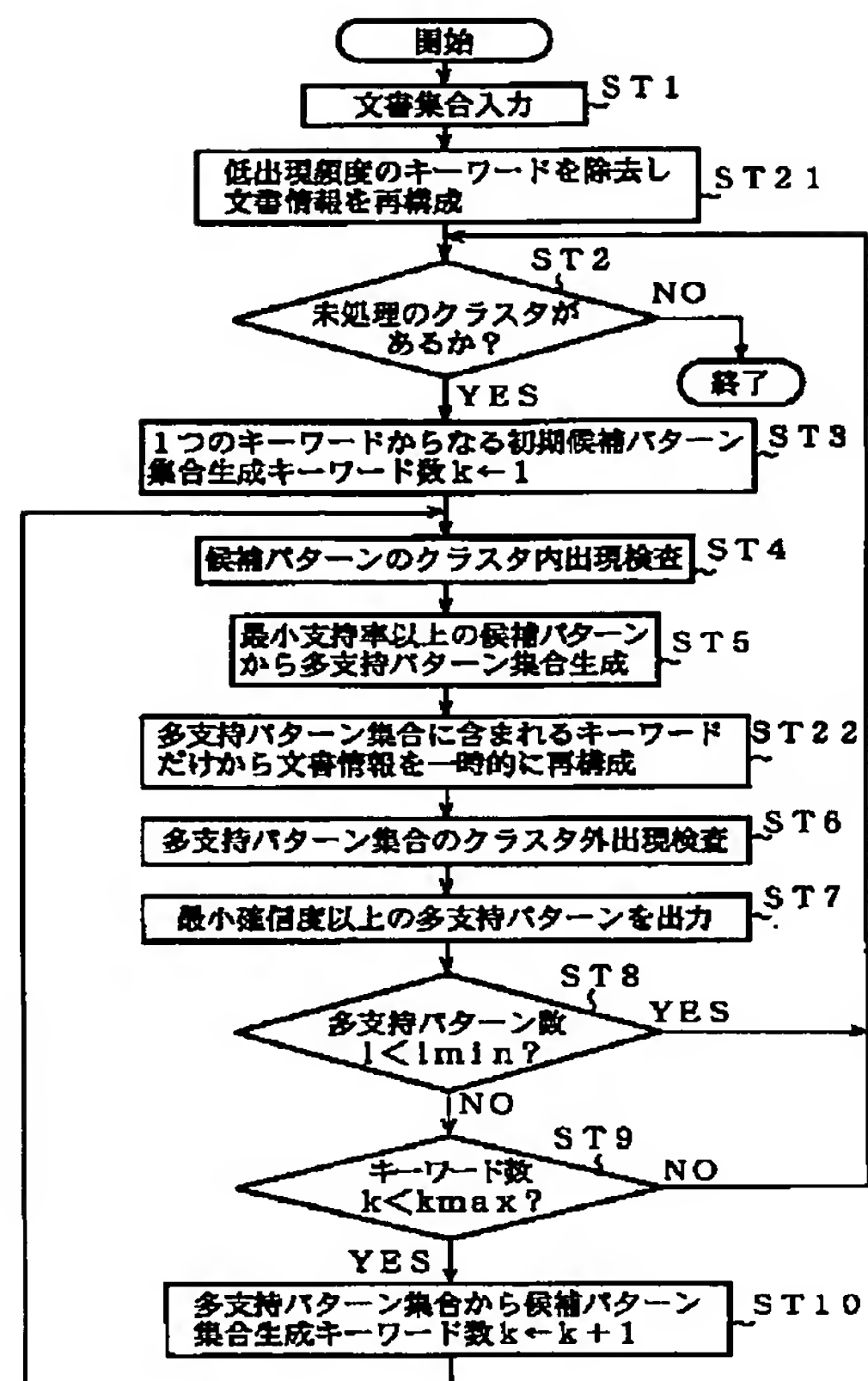
4：多支持パターン集合生成部（第1の選択手段、選択手段）  
5：高確信度パターン出力部（第2の選択手段、出力手段）  
6：候補パターン集合生成部（初期パターン集合生成手段、パターン集合生成手段）



【図2】



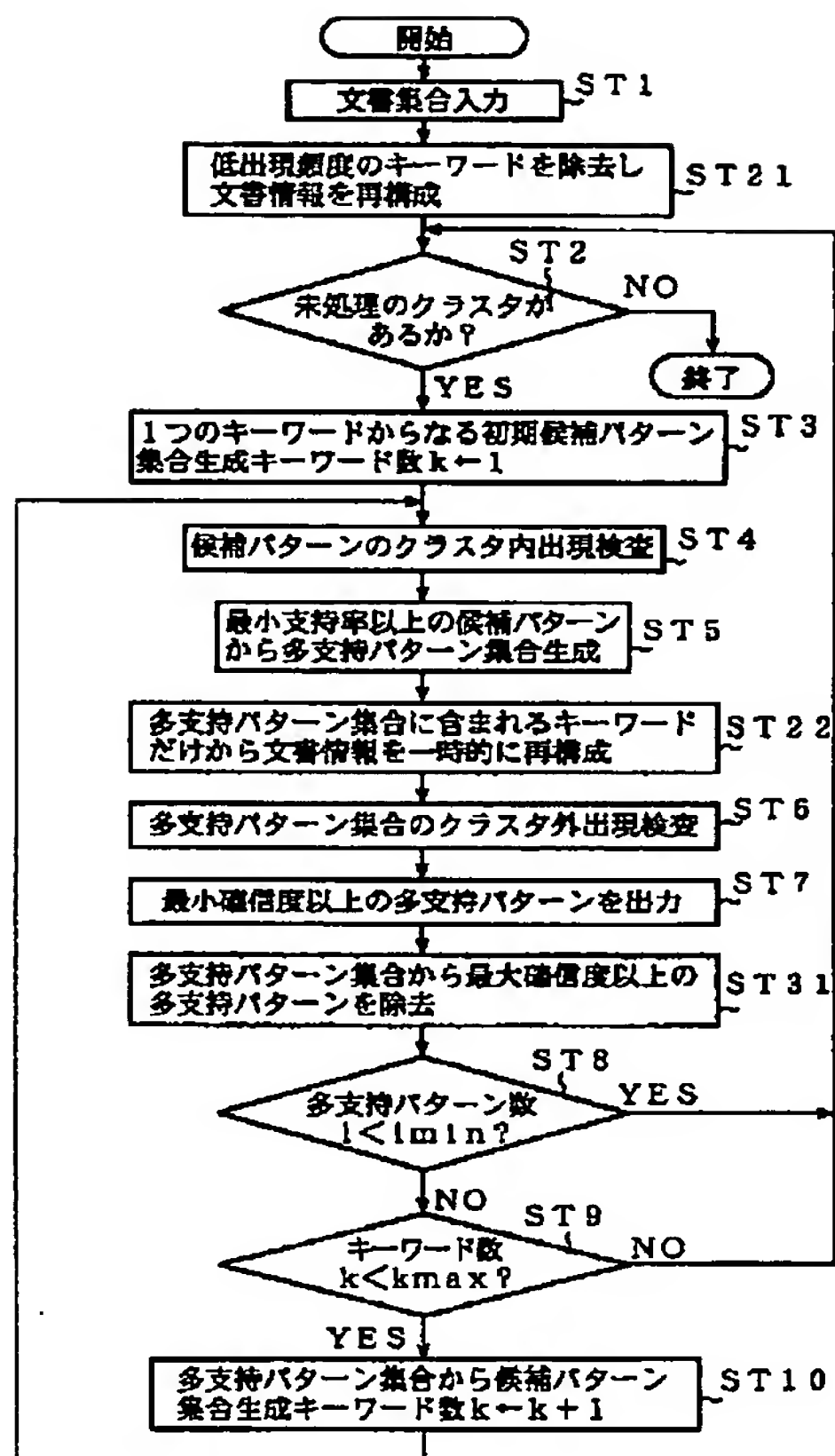
【図3】



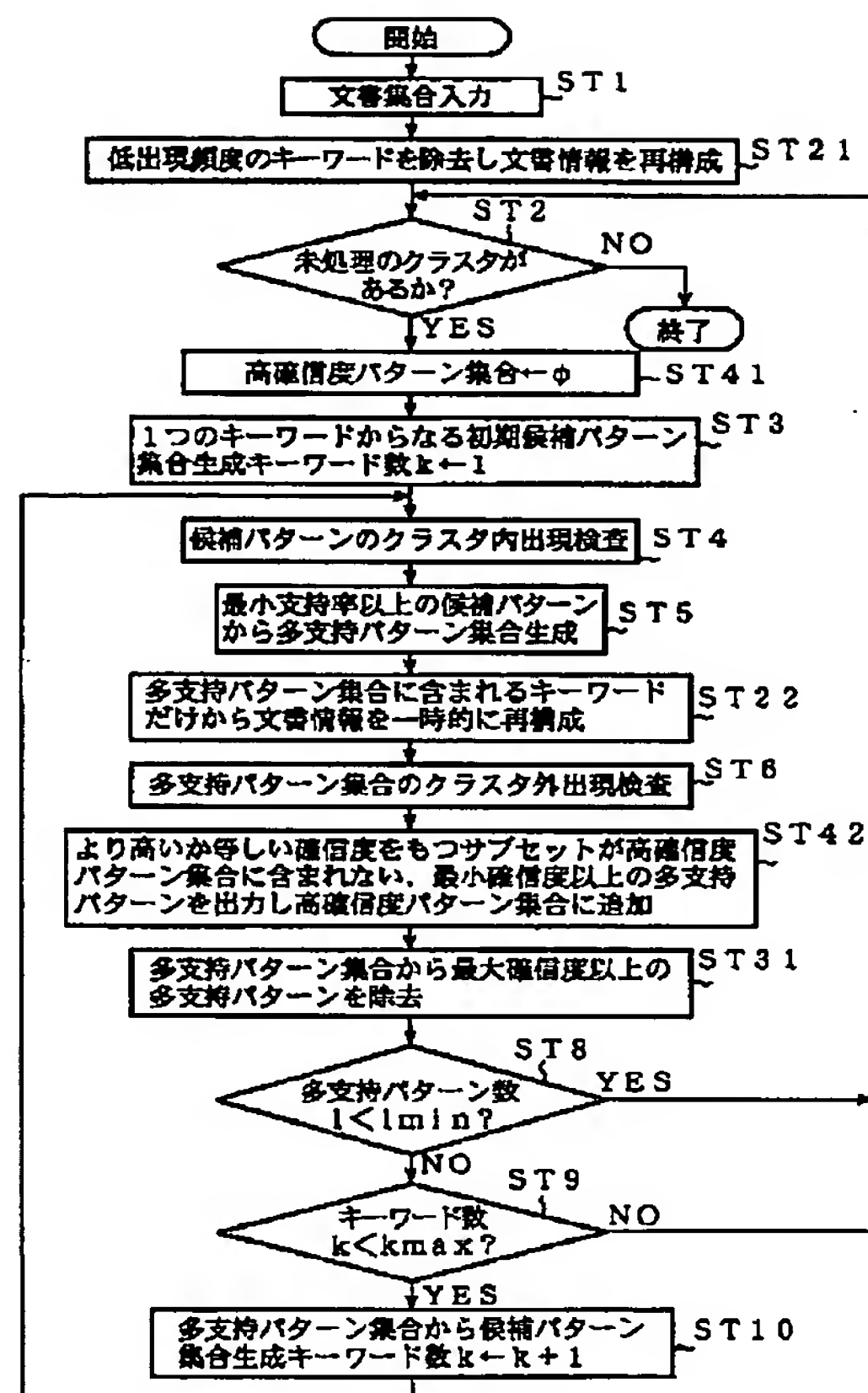
【図9】

文書ID		キーワード	
ファイル名			内容
高校サッカー	記事11361	高校サッカー	ゴール A高校 B高校 得点者 後半...
	記事11357	高校サッカー	C高校 D高校 PK 審判...
	記事11511	高校サッカー	A高校 C高校 延長 優勝...
Jリーグ	記事13131	サッカー	Jリーグ Eチーム Fチーム 交替 フリーキック...
	記事13135	サッカー	Jリーグ ドリブル Gチーム Hチーム Vゴール...
高校総体総合	記事15111	高校総体 総合	I選手 つり輪 あん馬 得点...
	記事15115	高校総体 総合	J選手 優勝 平行棒 跳馬 得点 逆転...
高校総体つり輪	記事15311	高校総体 種目別	つり輪 I選手 得点 難度...
	記事15335	高校総体 種目別	つり輪 K選手 落下 得点...
⋮			

【図4】

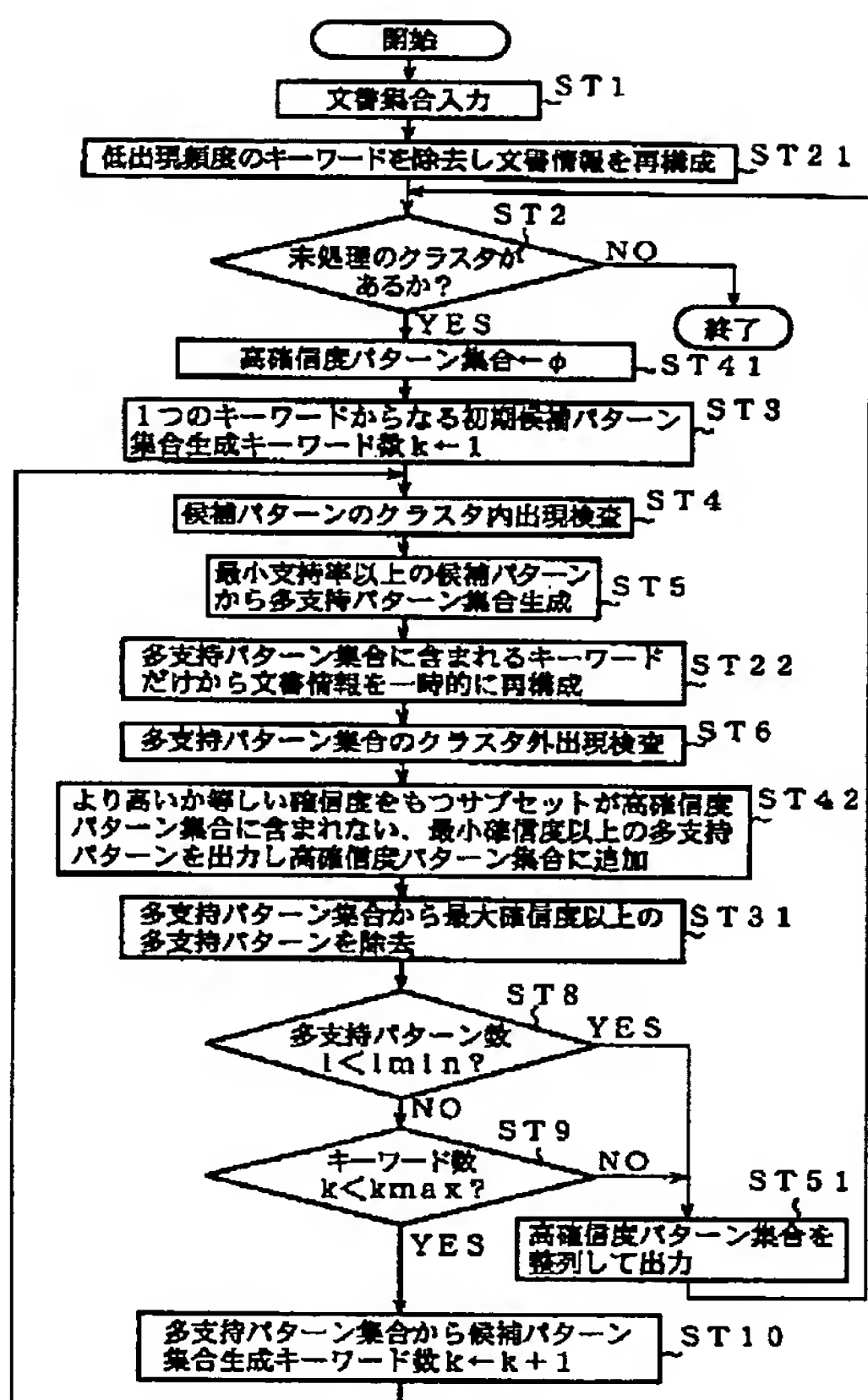


【図5】

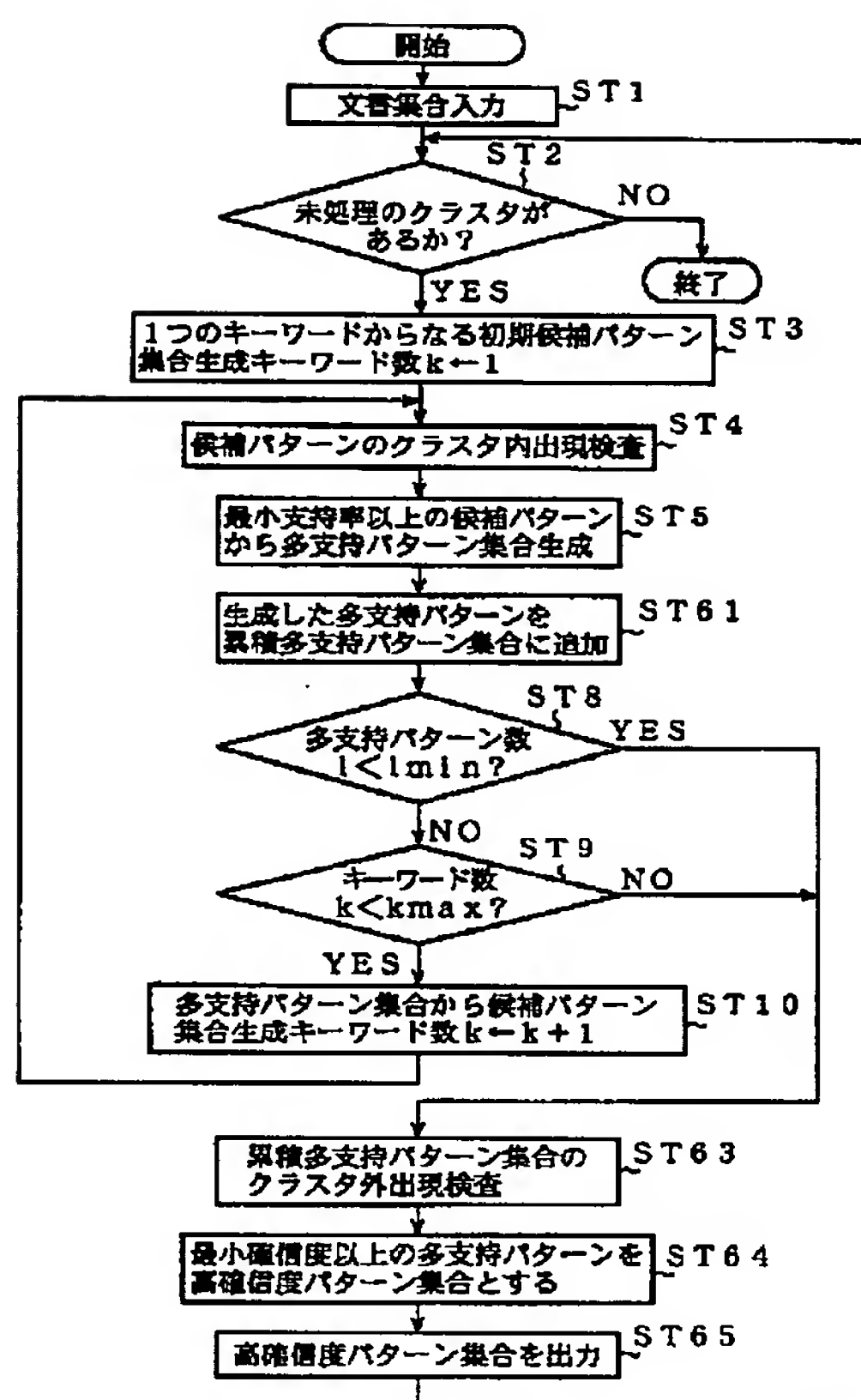




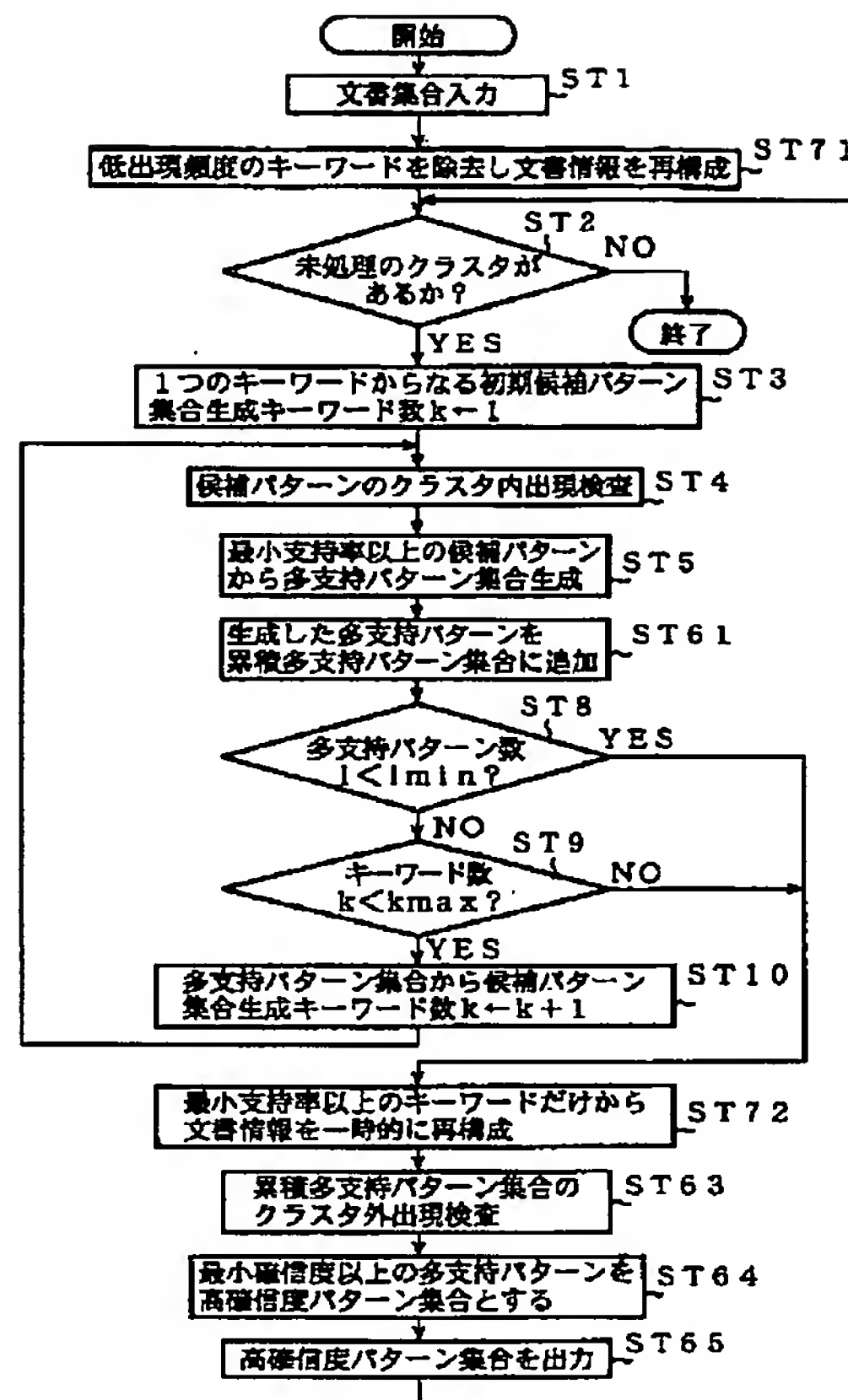
【図6】



【図7】



【図8】





(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平11-73415

(43)公開日 平成11年(1999) 3 月16日

(51)Int.Cl.<sup>9</sup>

識別記号

F I

G 0 6 F 17/30

G 0 6 F 15/403

3 5 0 C

審査請求 未請求 請求項の数4 O L (全 11 頁)

(21)出願番号 特願平9-231363

(22)出願日 平成9年(1997) 8 月27日

(71)出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(71)出願人 000221052

東芝コンピュータエンジニアリング株式会  
社

東京都青梅市新町3丁目3番地の1

(72)発明者 田野崎 康雄

東京都青梅市末広町2丁目9番地 株式会  
社東芝青梅工場内

(74)代理人 弁理士 須山 佐一

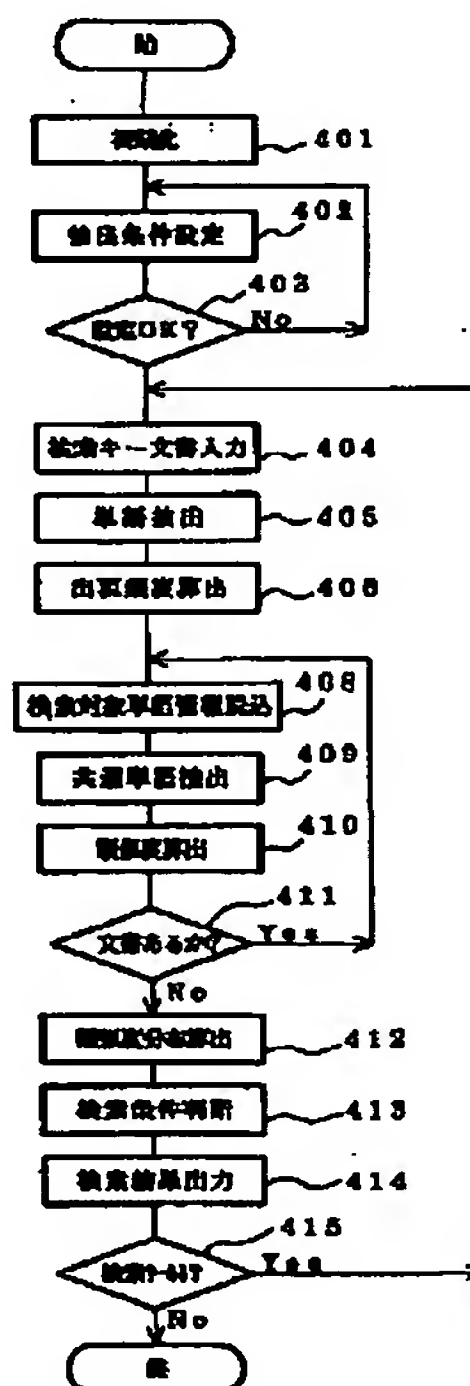
最終頁に続く

(54)【発明の名称】 類似文書検索装置及び類似文書検索方法

(57)【要約】

【課題】 ある文書（検索キー文書）と類似する文書を複数の検索対象文書のなかから検索する装置において、より信憑性の高い類似文書検索を実現する。

【解決手段】 検索キー文書と各検索対象文書との各々の類似度値の統計分布（例えば類似度の平均値）を求め、この統計分布を基準に、ユーザが設定した条件を満足するものを類似文書として抽出する。従来のように単に類似度値が高いものを類似文書として抽出する方式に比べ、類似文書としてより信憑性の高いものを検索結果として得ることができる。また、検索キー文書と各検索対象文書との類似度がどれも一般的な評価基準において高いとは言えないような場合に、類似文書がないことを検索結果として出力する。



## 【特許請求の範囲】

【請求項 1】 検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、

前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、

前記統計情報を基準とする類似文書の抽出条件を設定する抽出条件設定手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度および前記抽出条件設定手段により設定された抽出条件に基づいて類似文書を検索する検索手段とを具備することを特徴とする類似文書検索装置。

【請求項 2】 検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、

前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、

前記統計情報を基準とする類似文書の有無の判定条件を設定する判定条件設定手段と、

前記類似度算出手段によって算出された各検索対象文書の類似度および前記判定条件設定手段により設定された判定条件に基づいて類似文書の有無を判定する判定手段とを具備することを特徴とする類似文書検索装置。

【請求項 3】 検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、

前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、

前記算出された各検索対象文書の類似度の統計情報を求める工程と、

前記統計情報を基準とする類似文書の抽出条件を設定する工程と、

前記算出された各検索対象文書の類似度および前記設定された抽出条件に基づいて類似文書を検索する工程とを具備することを特徴とする類似文書検索方法。

【請求項 4】 検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、

前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、

前記算出された各検索対象文書の類似度の統計情報を求める工程と、

前記統計情報を基準とする類似文書の有無の判定条件を設定する工程と、

前記算出された各検索対象文書の類似度および前記設定された判定条件に基づいて類似文書の有無を判定する工

程とを具備することを特徴とする類似文書検索方法。

## 【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、電子化された文書データの検索装置に係り、特にある文書データを検索キーとしてこれと類似した文書データを自動検索する類似文書検索装置および類似文書検索方法に関する。

【0002】

【従来の技術】近年、大量の電子化された文書データが流通するようになり、自動分類等を行う目的で、文書データベース中から指定された文書（以下、検索キー文書と呼ぶ。）に類似する文書の自動検索を行うシステムが実用されてきている。従来の類似文書検索システムでは、検索キー文書に含まれている単語と他の文書（以下、検索対象文書と呼ぶ。）に含まれている単語とを比較し、共通する単語の種類や出現回数・場所などからベクトル空間法により類似度を算出し、最も類似度の高い検索対象文書を検索結果として出力したり、類似度の高い文書から順に出力していた。

【0003】

【発明が解決しようとする課題】従来の類似文書検索方式は、検索キー文書と文書データベース中の各検索対象文書との類似度を各々算出し、より類似度の高い文書を判定する、例えば最大類似度のものを検索結果として出力している。しかし、その検索結果は必ずしも妥当なものとは言えない。すなわち、上記従来の類似文書検索方式により得た類似度に基づく検索結果は、各検索対象文書に対して求めた各類似度の単純な大小比較により得た検索結果にすぎないため、例えば、検索キー文書と特に類似している検索対象文書が 1 つも存在しないような場合でも、その中で最も類似度の高い検索対象文書を類似文書として無条件に出力してしまう。

【0004】本発明はこのような課題を解決するためのもので、複数の検索対象文書のなかから検索キー文書との類似が際立っているものを確実に検索することのできる類似文書検索装置および類似文書検索方法の提供を目的としている。

【0005】また、本発明は、多くの一般的な類似評価の基準において類似していると呼べる類似文書のみを検索結果として得ることで、信頼性の向上を図ることのできる類似文書検索装置および類似文書検索方法の提供を目的としている。

【0006】

【課題を解決するための手段】上記目的を達成するために、本発明の類似文書検索装置は、請求項 1 に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める



統計情報算出手段と、前記統計情報を基準とする類似文書の抽出条件を設定する抽出条件設定手段と、前記類似度算出手段によって算出された各検索対象文書の類似度および前記抽出条件設定手段により設定された抽出条件に基づいて類似文書を検索する検索手段とを具備することを特徴とする。

【0007】また、本発明の類似文書検索装置は、請求項2に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、前記統計情報を基準とする類似文書の有無の判定条件を設定する判定条件設定手段と、前記類似度算出手段によって算出された各検索対象文書の類似度および前記判定条件設定手段により設定された判定条件に基づいて類似文書の有無を判定する判定手段とを具備することを特徴とする。

【0008】さらに、本発明の類似文書検索方法は、請求項3に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、前記算出された各検索対象文書の類似度の統計情報を求める工程と、前記統計情報を基準とする類似文書の抽出条件を設定する工程と、前記算出された各検索対象文書の類似度および前記設定された抽出条件に基づいて類似文書を検索する工程とを具備することを特徴とする。

【0009】さらに、本発明の類似文書検索方法は、請求項4に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、前記算出された各検索対象文書の類似度の統計情報を求める工程と、前記統計情報を基準とする類似文書の有無の判定条件を設定する工程と、前記算出された各検索対象文書の類似度および前記設定された判定条件に基づいて類似文書の有無を判定する工程とを具備することを特徴とする。

【0010】請求項1および請求項3の発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の抽出条件に基づいて類似文書を検索することで、類似文書としてより妥当性の高いもの、つまり類似度がその他の多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として検索することができる。

【0011】請求項2および請求項4の発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の有無の判定条件に基づいて類似文書の有無を判定することで、検索キー文書と各検索対象文書との類似度が

ずれも多く一般的な評価基準において高いと言えないような場合に類似文書が存在しないと、一般的な評価基準において類似していると言える類似文書だけを検索結果として得ることができる。

【0012】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して詳細に説明する。

【0013】図1は本発明に係る一実施形態の類似文書検索装置のハードウェア構成を示す図である。

10 【0014】同図に示すように、この類似文書検索装置は、CPUおよびメモリなどから構成される制御装置1、キーボードなどの入力装置2、類似文書の検索結果などを表示する表示装置3、および文書データや類似文書検索のための各文書の単語情報などを格納する外部記憶装置4から構成される。

【0015】図2に本類似文書検索装置における制御装置1の構成を示す。制御装置1は制御部200とメモリ部229からなる。

20 【0016】制御部200は、初期化部201、入力部202、出力部203、検索対象文書読み出し部204、検索対象単語抽出部205、検索対象単語出現頻度算出部206、検索対象単語情報書込部207、検索キー文書入力部208、検索キー単語抽出部209、検索キー単語出現頻度算出部210、検索対象単語情報読み出し部211、共通単語抽出部212、類似度算出部213、類似度統計分布計算部214、抽出条件設定部215、検索結果出力部216などから構成される。メモリ部229は、検索対象文書格納バッファ部230、検索対象単語情報格納バッファ部231、検索キー文書格納バッファ部232、検索キー単語情報格納バッファ部234、共通単語情報格納バッファ部235、類似度格納バッファ部236、抽出条件設定バッファ部237、類似度統計分布結果バッファ部238、検索結果出力バッファ部239などから構成される。

30 【0017】初期化部201は、上記各バッファ部の初期化を行う。入力部202は、ユーザによる入力装置2からの検索キー文書や抽出条件の設定など各種設定の入力を行う。出力部203は、入力部202により入力された検索キー文書などの各種設定内容を表示装置3に出力する。

40 【0018】検索対象文書読み出し部204は、外部記憶装置4に格納されている検索対象文書に関する情報を文書データベース化するために、文書データベース化すべき文書情報を外部記憶装置4から読み込み、検索対象文書格納バッファ部230に格納する。

50 【0019】検索対象単語抽出部205は、検索対象文書格納バッファ部230に格納されている検索対象文書からの単語の切り出しを行う。そして、切り出した単語のなかからその文書内容を表す上でキーとなる単語を抽出し、抽出した単語種を検索対象単語情報格納バッファ

部231に格納する。単語の切り出しは形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【0020】検索対象単語出現頻度算出部206は、検索対象単語抽出部205により抽出された個々のキー単語について、検索対象文書中での出現頻度を算出し、これを検索対象文書の単語情報として検索対象単語情報格納バッファ部231に格納する。

【0021】検索対象単語情報書込部207は、検索対象単語情報格納バッファ部231に格納されている検索対象文書の単語情報を外部記憶装置4に格納する。

【0022】検索キー文書入力部208は、入力装置2から入力された検索キー文書の情報を検索キー文書格納バッファ部232に格納する。

【0023】検索キー単語抽出部209は、検索キー文書格納バッファ部232に格納されている検索キー文書からの単語切り出しを行う。そして、その文書の内容を表す上でキーとなる単語を抽出し、抽出した単語種を検索キー単語情報格納バッファ部234に格納する。単語の切り出しは形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【0024】検索キー単語出現頻度算出部210は、検索キー単語抽出部209により抽出された個々のキー単語について、検索キー文書中での出現頻度を算出し、これを検索キー文書の単語情報として検索キー単語情報格納バッファ部234に格納する。

【0025】検索対象単語情報読み出し部211は、外部記憶装置4に格納されいてる各検索対象文書の単語情報（単語の出現頻度情報）を1文書分ごとに呼び出し、検索対象単語情報格納バッファ部231に格納する。

【0026】共通単語抽出部212は、検索キー単語情報格納バッファ部234に格納されている検索キー文書の単語情報と検索対象単語情報格納バッファ部231に格納されている検索対象文書の単語情報とを比較して、一致する単語の種類と出現頻度情報を共通単語情報格納バッファ部235に格納する。

【0027】類似度算出部213は、共通単語情報格納バッファ部235に格納されている情報に基づき検索キー文書と検索対象文書との類似度を算出し、その類似度値を類似度格納バッファ部236に格納する。

【0028】類似度統計分布計算部214は、類似度格納バッファ部236に格納されている検索キー文書と全検索対象文書との類似度値から類似度の平均値や標準偏差値などの統計分布情報を求めて類似度統計分布結果バッファ部238に格納する。抽出条件設定部215は、入力装置2を介してユーザより入力された、類似度統計結果から類似文書の検索結果を抽出する場合の条件、または類似度統計結果から検索キー文書との類似文書があ

るとするための条件、または検索キー文書との類似文書がないとするための条件などの抽出条件値を抽出条件設定バッファ部237に格納（設定）する。

【0029】検索結果出力部216は、類似度統計分布結果バッファ部238に格納されている統計分布情報、抽出条件設定バッファ部237に格納されている抽出条件値、さらには類似度格納バッファ部236に格納されている各検索対象文書の類似度値から、検索キー文書に対する類似文書検索結果として、検索対象文書の有無、検索対象文書が有る場合の該当文書を判断し、その検索結果を検索結果出力バッファ部239に格納し、そして検索結果出力バッファ部239の内容を表示装置3に出力する。

【0030】次に、本実施形態の類似文書検索装置の動作を説明する。

【0031】最初に、検索対象文書のデータベースの作成手順を図3、図5、図6により説明する。図3はその手順を示すフローチャートである。

【0032】まず、初期化部201により全バッファ部の初期化を行う（ステップ301）。続いて検索対象文書読み出し部204が、外部記憶装置4から複数のテキスト文書を読み出し、検索対象文書格納バッファ部230に検索対象文書として格納する（ステップ302）。具体例として、例えば図5に示すような内容のテキスト文書を検索対象文書の一つとして格納したとする。

【0033】次に、検索対象単語抽出部205が、検索対象文書格納バッファ部230に格納されている個々の検索対象文書について、形態素解析などによって単語の切り出しを行い、切り出した単語のなかから文書内容を表すキー単語を抽出し、そのキー単語の単語種（例えば品詞情報）を検索対象単語情報格納バッファ部231に格納する（ステップ303）。

【0034】次に、検索対象単語出現頻度算出部206が、検索対象単語情報格納バッファ部231に格納されている検索対象文書のキー単語について、検索対象文書全体での出現頻度を算出し、その結果を検索対象単語情報格納バッファ部231に格納する（ステップ304）。図6に検索対象単語情報格納バッファ部231の格納例を示す。このバッファ部231において単語と頻度は対応付けて記述される。例えばキー単語「文書」が文書全体のなかで2回出現している場合は頻度として「2」が記述される。

【0035】このようにして検索対象単語情報格納バッファ部231に格納された情報は、検索対象文書のデータベースとして外部記憶装置4に蓄積される（ステップ305）。

【0036】この後、検索対象文書格納バッファ部230に文書データベース化前の検索対象文書が残っているかどうかを判断し（ステップ306）、他に検索対象文書があればステップ302に戻って、その新たな検索対

10

20

30

40

50



象文書についての前記同様の文書データベースの作成が行われる。他に検索対象文書がなければ本処理を終了する。

【0037】次に、類似文書の検索手順を図4、図7乃至図16により説明する。図4は類似文書検索手順を示すフローチャートである。

【0038】まず、初期化部201により全バッファ部の初期化を行う（ステップ401）。続いて抽出条件設定部215が起動される。抽出条件設定部215は、入力装置2を通じてユーザより、

1. 類似度統計結果から類似文書の検索結果を抽出する場合の条件
2. 類似度統計結果から類似文書が存在しているとするための条件
3. 類似度統計結果から類似文書が存在していないとするための条件

などの抽出条件値の入力を受け付けて抽出条件設定バッファ部237に格納（設定）する（ステップ402）。より具体的には、図7に抽出条件設定バッファ部237の格納例を示しているように、1. の条件として、「平均類似度の2倍以上の類似度を持つ検索対象文書を検索結果とする。」などが設定される。

【0039】2. の条件として、「平均類似度の2倍以上の類似度を持つ検索対象文書がある場合」などが設定される。

【0040】3. の条件として、「すべての検索対象文書の類似度が0.1以下である場合」などが設定される。

【0041】これらの抽出条件値はユーザにより任意に決定される。

【0042】この抽出条件の設定が完了したら（ステップ403）、検索キー文書入力部208が起動される。検索キー文書入力部208は、入力装置2を通じてユーザより検索キー文書の入力を受け付け、入力した検索キー文書の情報を検索キー文書格納バッファ部232に格納する（ステップ404）。具体例として、図8に示すような検索キー文書が入力されてバッファ部232に格納されたとする。

【0043】次に、検索キー単語抽出部209が、検索キー文書格納バッファ部232に格納されている検索キー文書から形態素解析などによって単語の切り出しを行い、切り出した単語のなかから文書内容を表すキー単語を抽出し、そのキー単語の単語種（例えば品詞情報）を検索キー単語情報格納バッファ部234に格納する（ステップ405）。

【0044】次に、検索キー単語出現頻度算出部210が、検索キー単語情報格納バッファ部234に格納されている個々のキー単語について、検索キー文書全体の中での出現頻度を算出し、これを検索キー単語情報格納バッファ部234にキー単語と対応付けて格納する（ス

テップ406）。図9に検索キー単語情報格納バッファ部234の格納例を示す。このバッファ部234においてキー単語と頻度は対応付けて記述され、例えばキー単語「今後」が2回出現している場合は頻度として「2」が記述される。

【0045】次に、検索対象単語情報読み出し部211が、外部記憶装置4に格納されている各検索対象文書の単語情報を1文書ごとに読み込み、検索対象単語情報格納バッファ部231に書き込む（ステップ408）。

- 10 【0046】この後、共通単語抽出部212が起動され、共通単語抽出部212は、検索対象単語情報格納バッファ部231と検索キー単語情報格納バッファ部234とに共通に格納されているキー単語を検出し、共通単語情報格納バッファ部235に格納する（ステップ409）。例えば、図10に示すように、図6の検索対象単語情報格納バッファ部231と図9の検索キー単語情報格納バッファ部234に共通するキー単語として「画像」が検出され、このキー単語「画像」とその頻度情報「3」を共通単語情報格納バッファ部235に対応付けて格納する。
- 20

【0047】次に、類似度算出部213が、共通単語情報格納バッファ部235に格納されている情報に基づき検索キー文書と検索対象文書との類似度をベクトル空間法などにより算出し、その類似度値を類似度格納バッファ部236に格納する（ステップ410）。例えば、図11に示すような各検索対象文書ごとの類似度が類似度格納バッファ部236に格納される。

- 30 【0048】すべての検索対象文書について類似度計算が完了すると（ステップ411）、類似度統計分布計算部214が起動する。類似度統計分布計算部214は、類似度格納バッファ部236に格納されている各検索対象文書の類似度の統計分布を算出し、その結果を類似度統計分布結果バッファ部238に格納する（ステップ412）。例えば、図12に示すように、各検索対象文書の類似度の平均値「0.25」を類似度の統計分布情報として求める。

- 40 【0049】次に、検索結果出力部216が、抽出条件設定バッファ部237に格納されている抽出条件値を判断し（ステップ413）、その抽出条件値、類似度統計分布結果バッファ部238に格納されている統計分布情報、さらには類似度格納バッファ部236に格納されている各検索対象文書の類似度値から、検索キー文書に対する類似文書の検索結果として、検索対象文書の有無や、検索対象文書が有る場合の該当文書を判断し、その結果を検索結果出力バッファ部239に格納する。例えば、1. の条件（類似度統計結果から類似文書の検索結果を抽出する場合の条件）に対し、図13に示すように、上記条件を満足する検索対象文書があれば、そのIDを検索結果出力バッファ部239に格納する。

- 50 【0050】また、2. の条件（類似度統計結果から検

索キー文書との類似文書が存在しているとするための条件) に対し、該条件を満足している場合は、図14に示すように、当該検索キー文書を類似文書有りの検索キー文書として、そのIDを検索結果出力バッファ部239に格納する。

【0051】さらに、3. の条件(類似度統計結果から検索キー文書との類似文書が存在していないとするための条件) に対し、該条件を満足する場合は、図15に示すように、当該検索キー文書を類似文書無しの検索キー文書として、そのIDを検索結果出力バッファ部239に格納する。

【0052】検索結果出力部216は、検索結果出力バッファ部239の内容を、例えば図16に示すような形式で表示装置3に出力する(ステップ414)。図16の例では、例えば図14に示すID「2」の類似文書有りの検索キー文書と類似する検索対象文書としてID「1」「42」「54」「314」などがあることを示している。また、2. と3. の各条件により、類似する検索対象文書がないと判断された検索キー文書に対しては類似文書がないことが表示される。

【0053】この後、次の検索キー文書がある場合はステップ404に戻ってその検索キー文書を入力し、以降同様の処理を行う。次の検索キー文書がなければ本処理を終了する。

【0054】なお、本動作例では、1つの検索キー文書に対する検索結果を得たところでこれを表示するようにしたが、すべての検索キー文書に対する検索結果を得た後、検索結果を見たい検索キー文書を指定すると、その検索結果が表示されるようにしてもよい。

【0055】このように本実施形態の類似文書検索装置においては、検索キー文書と各検索対象文書との各々の類似度値の統計分布(この実施形態では類似度の平均値)を求め、この統計分布を基準に、ユーザが設定した条件を満足するものを類似文書として抽出することで、従来のように単に類似度値が高いものを類似文書として抽出する方式に比べ、類似文書としてより信憑性の高いものを検索結果として得ることができる。すなわち、本実施形態は、類似度の統計分布を基準としているので、検索キー文書との類似度がその他多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として得られる。

【0056】また、従来方式では、検索キー文書と各検索対象文書との類似度が、どれも一般的な評価基準において高いとは言えないような場合でも検索結果として類似文書が無条件に出力してしまうが、本実施形態では、このような場合において類似文書がないことを検索結果として出力する。このような点からも、本実施形態の類似文書検索装置によれば、類似文書として信憑性の高い検索結果を得ることができ、類似文書検索効率を大幅に高めることができる。

【0057】

【発明の効果】以上説明したように本発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の抽出条件に基づいて類似文書を検索することで、類似文書としてより妥当性の高いもの、つまり類似度がその他の多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として得ることができる。

【0058】また、本発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の有無の判定条件に基づいて類似文書の有無を判定することで、検索キー文書と各検索対象文書との類似度がいずれも一般的な評価基準において高いと言えないような場合に類似文書が存在しないとし、一般的な評価基準において類似していると言える類似文書だけを検索結果として得ることができる。

【図面の簡単な説明】

【図1】本発明に係る一実施形態の類似文書検索装置のハードウェア構成を示す図

【図2】図1の類似文書検索装置における制御装置の機能ブロック図

【図3】検索対象文書のデータベースの作成手順を示す図

【図4】類似文書検索手順を示す図

【図5】検索対象文書の例

【図6】検索対象単語情報の格納例

【図7】抽出条件の設定例

【図8】検索キー文書の例

【図9】検索キー単語情報の例

【図10】共通単語情報の例

【図11】各検索対象文書の類似度の例

【図12】類似度の平均値の例

【図13】類似文書の検索結果の例

【図14】類似文書の検索結果の例

【図15】類似文書の検索結果の例

【図16】類似文書の検索結果の出力例

【符号の説明】

200……制御部

210……初期化部

202……入力部

203……出力部

204……検索対象文書読み出し部

205……検索対象単語抽出部

206……検索対象単語出現頻度算出部

207……検索対象単語情報書込部

208……検索キー文書入力部

209……検索キー単語抽出部

210……検索キー単語出現頻度算出部

211……検索対象単語情報読み出し部

11

12

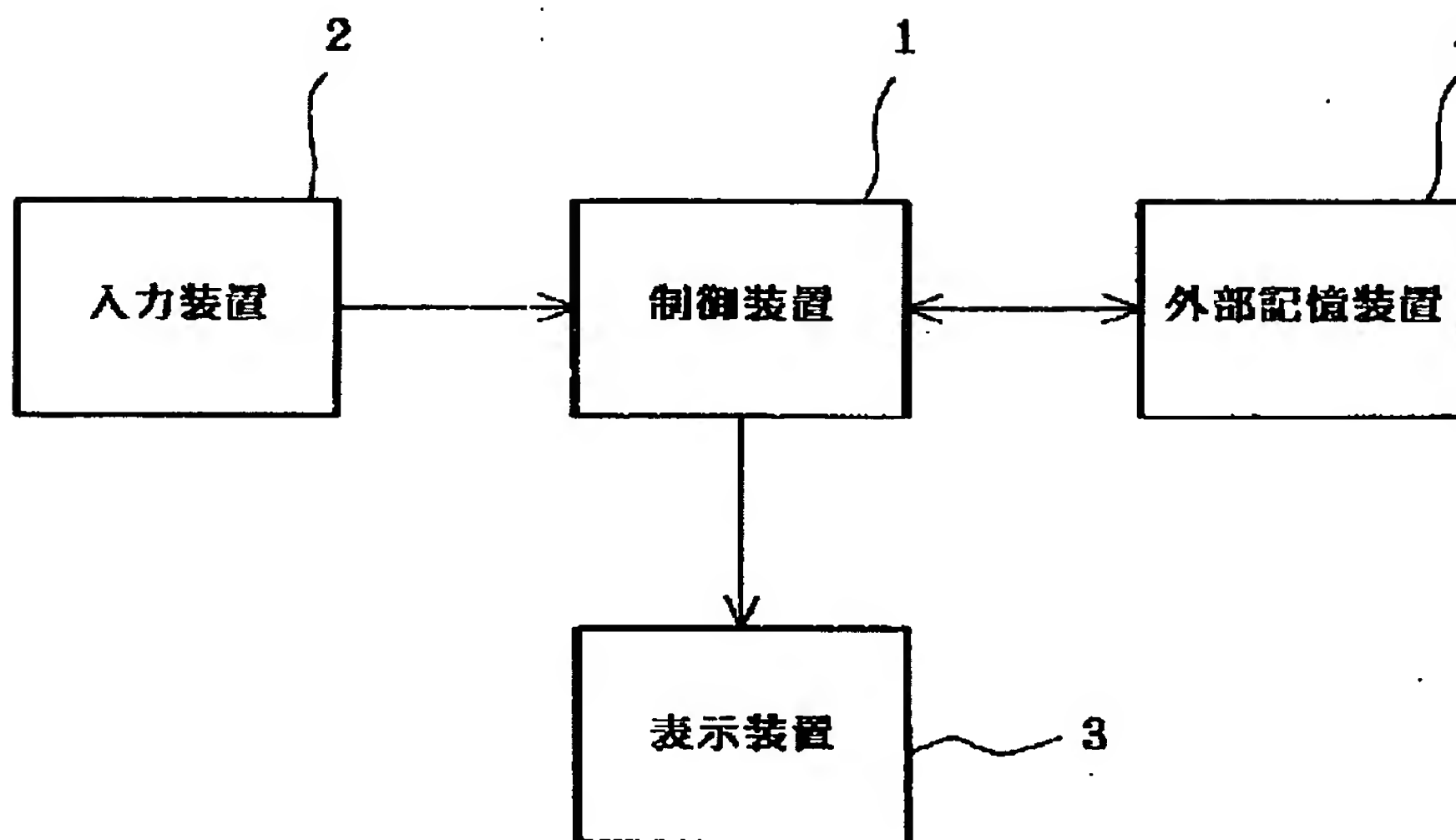
212……共通単語抽出部  
 213……類似度算出部  
 214……類似度統計分布計算部  
 215……抽出条件設定部  
 216……検索結果出力部  
 229……メモリ部  
 230……検索対象文書格納バッファ部  
 231……検索対象単語情報格納バッファ部

\*

\*232……検索キー文書格納バッファ部  
 233……検索キー単語情報格納バッファ部  
 235……共通単語情報格納バッファ部  
 236……類似度格納バッファ部  
 237……抽出条件設定バッファ部  
 238……類似度統計分布結果バッファ部  
 239……検索結果出力バッファ部

【図1】

【図5】



この文書は、画像について書いてあります。  
 ……  
 検索対象文書バッファ格納例

【図8】

今後は、画像に関する事項が…  
 検索キー文書バッファ格納例

【図6】

【図7】

単語	類似度
文書	2
画像	3
……	……

検索対象単語情報バッファ格納例

抽出検索対象文書＝ 平均類似度の2倍以上
類似検索対象文書有り＝ 平均類似度の2倍以上の文書がある
類似度検索対象文書無し＝ すべての検索対象文書の類似度が0.1以下

抽出条件設定バッファ格納例

【図9】

【図10】

【図12】

単語	類似度
今後	2
画像	3
……	……

検索キー単語情報バッファ格納例

単語	類似度
画像	3
……	……

共通単語情報バッファ格納例

平均値＝0.25
----------

類似度統計分布結果バッファ格納例

【図14】

【図11】

【図13】

検索対象文書ID	類似度
1	0.2454
2	0.6842
3	0.9542
……	……

類似度バッファ格納例

類似検索対象文書＝ 1 42 54 314 ……
---

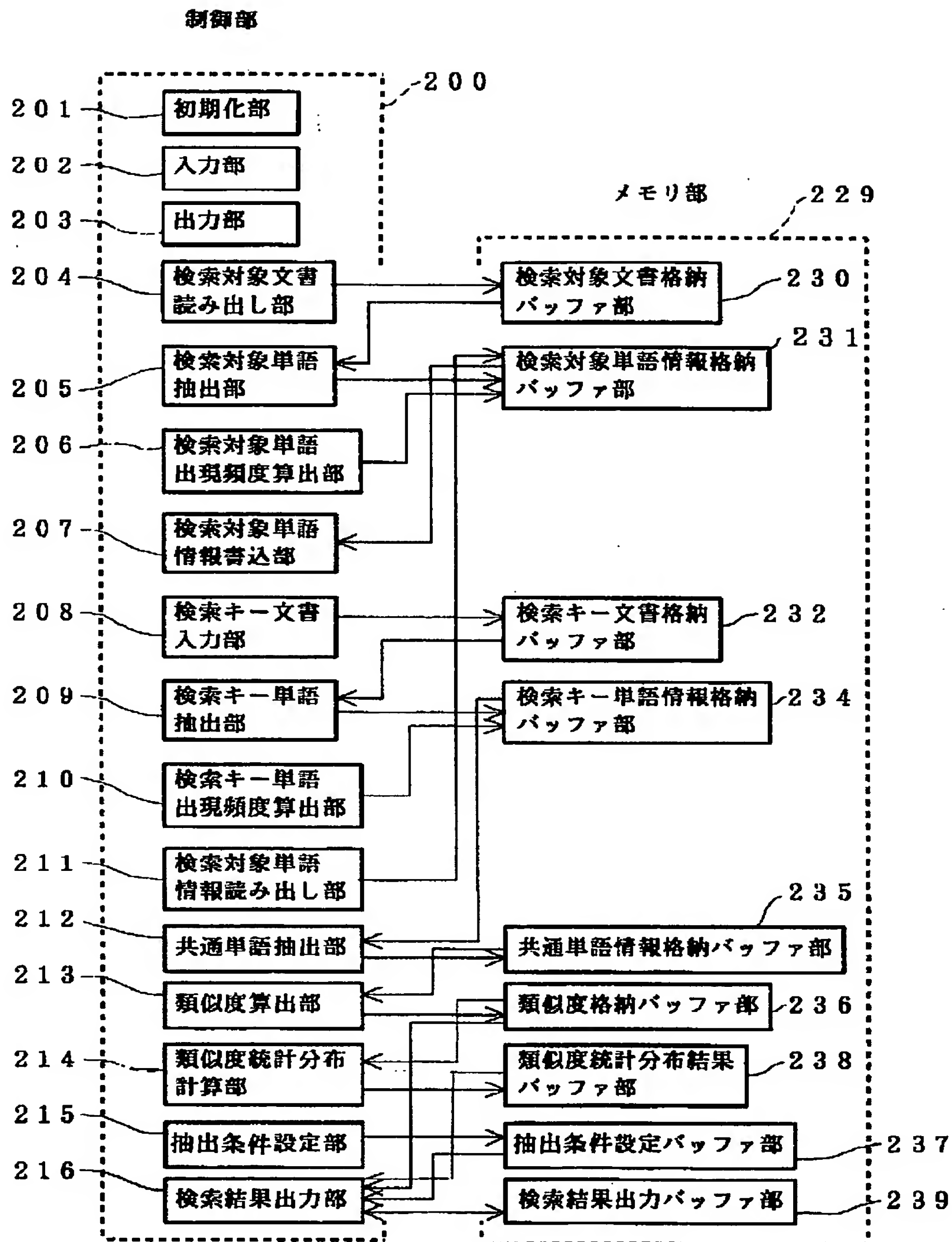
検索結果出力バッファ格納例

類似文書有り検索キー文書＝ 2 3 ……
-------------------------------

検索結果出力バッファ格納例



【図2】



【図15】

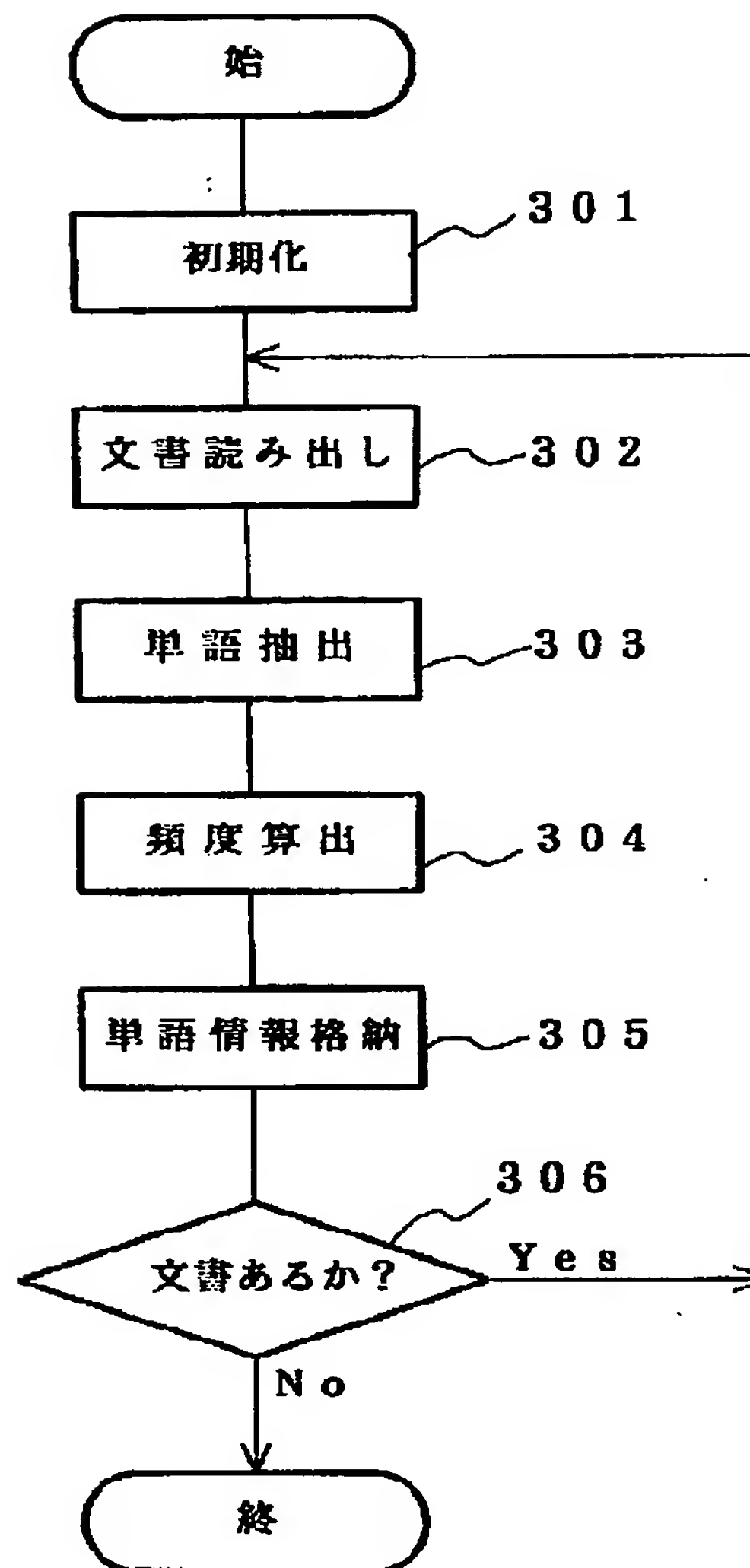
【図16】

類似文書無し検索キー文書＝  
4  
6  
.....

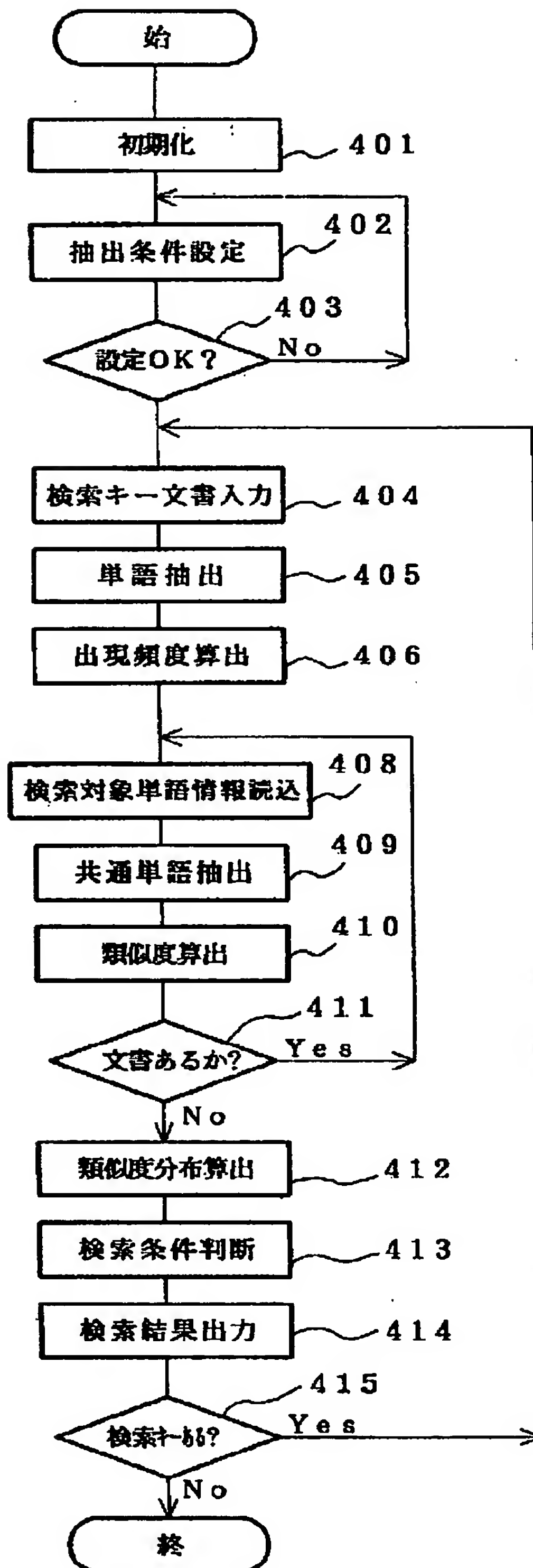
検索結果出力バッファ格納例

結果：類似検索対象文書  
1  
42  
54  
314

【図3】



【図4】





フロントページの続き

(72)発明者 中本 幸夫  
東京都青梅市新町1381番地 1 東芝コンピ  
ュータエンジニアリング株式会社内

(72)発明者 仁科 卓哉  
東京都青梅市新町1381番地 1 東芝コンピ  
ュータエンジニアリング株式会社内

(72)発明者 久保田 直秀  
東京都青梅市新町1381番地 1 東芝コンピ  
ュータエンジニアリング株式会社内

(19) 日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11) 特許出願公開番号

特開平11-345239

(43) 公開日 平成11年(1999)12月14日

(51) Int.Cl.<sup>5</sup>

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/401

15/40

3 1 0 A

3 7 0 A

審査請求 未請求 請求項の数11 O L (全 20 頁)

(21) 出願番号

特願平10-153231

(22) 出願日

平成10年(1998)6月2日

(71) 出願人

000004226

日本電信電話株式会社

東京都千代田区大手町二丁目3番1号

(72) 発明者

巖寺 俊哲

東京都新宿区西新宿三丁目19番2号 日本

電信電話株式会社内

(74) 代理人

弁理士 伊東 忠彦

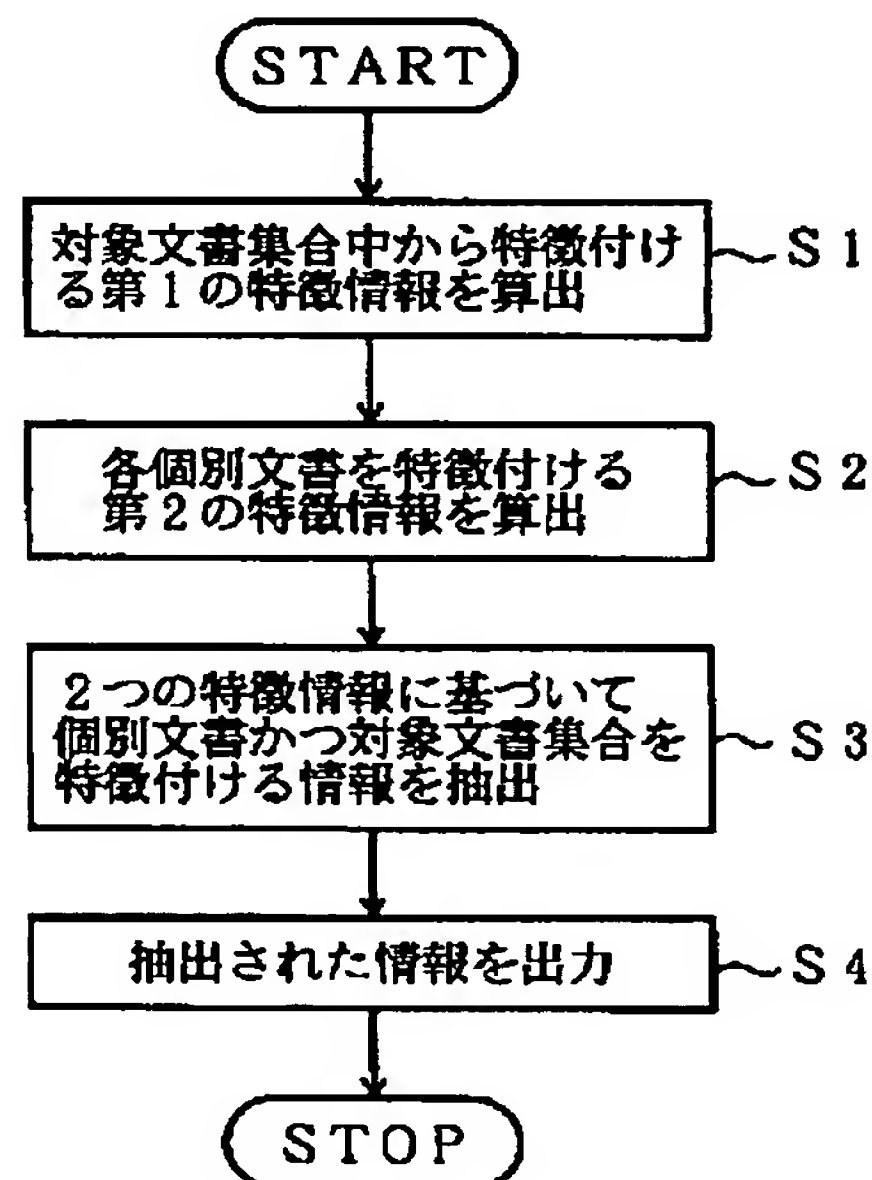
(54) 【発明の名称】 文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体

(57) 【要約】

【課題】 予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供する。

【解決手段】 本発明は、対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出し、対象文書集合中の各個別文書を他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出し、第1の特徴情報と、前記第2の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、抽出された前記情報を各個別文書の特徴づける情報として出力する。

本発明の原理を説明するための図



## 【特許請求の範囲】

【請求項 1】 文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出方法において、

前記対象文書集合を標準文書集合に対して特徴付ける第 1 の特徴情報を、該対象文書集合中から算出し、

前記対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第 2 の特徴情報を該対象文書集合中の各個別文書から算出し、

前記第 1 の特徴情報と、前記第 2 の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、

抽出された前記情報を各個別文書を特徴づける情報として出力することを特徴とする文書情報抽出方法。

【請求項 2】 文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出装置であって、

前記対象文書集合を標準文書集合に対して特徴付ける第 1 の特徴情報を、該対象文書集合中から算出する第 1 の特徴情報算出手段と、

前記対象文書集合中の各個別文書について、他の個別文書を特徴付ける第 2 の特徴情報を該対象文書集合中の各個別文書から算出する第 2 の特徴情報算出手段と、

前記第 1 の特徴情報算出手段で算出された前記第 1 の特徴情報と、前記第 2 の特徴情報算出手段で算出された前記第 2 の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出手段と、

抽出された前記情報を各個別文書を特徴づける情報として出力する特徴情報出力手段とを有することを特徴とする文書情報抽出装置。

【請求項 3】 前記入力記憶手段から標準文書集合を受け取る標準文書集合更新手段と、

前記標準文書集合更新手段に与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析手段と、

前記標準文書集合中の単語と該単語の出現頻度を対応付けて記憶する標準文書集合解析結果記憶手段と、

前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力手段と、

前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析手段と、

前記各対象文書集合中の単語と該単語の出現頻度を各文書を対応付けて記憶する対象文書集合解析結果記憶手段

と、

前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶する個別文書解析結果記憶手段と、

前記対象文書集合全体としての特徴情報を、前記対象文書集合解析結果記憶手段に記憶されている情報を用いて算出する対象文書集合全体特徴算出手段と、

前記対象文書集合全体特徴算出手段によって算出された、前記対象文書集合全体としての特徴情報を記憶する対象文書集合全体特徴記憶手段と、

10 前記対象文書集合中の各個別文書の特徴情報を、前記個別文書解析結果記憶手段に記憶されている情報を用いて算出する個別文書特徴算出手段と、

前記個別文書特徴算出手段によって算出された、前記対象文書集合中の各個別文書に対応する特徴情報を記憶する個別文書特徴記憶手段と、

前記個別文書解析結果記憶手段または、前記対象文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する目的情報一時記憶手段と、

20 前記対象文書集合解析結果記憶手段または、前記標準文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する基準情報一時記憶手段と、

前記目的情報一時記憶手段に記憶されているデータと前記基準情報一時記憶手段に記憶されているデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出手段と、

前記対象文書集合全体特徴記憶手段に記憶されているデータと前記個別文書特徴記憶手段に記憶されているデータを用いて、各個別文書の特徴情報を算出する特徴情報

30 算出手段と、

前記特徴情報算出手段において算出された各個別文書の特徴情報を記憶する特徴情報記憶手段と、

前記特徴情報記憶手段に記憶されている各個別文書の特徴情報を用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出手段と、

前記特徴表現抽出手段により前記各個別文書から抽出された特徴表現を記憶する特徴表現記憶手段と、

40 前記特徴表現記憶手段に記憶されている特徴表現を転送媒体に与える特徴表現出力手段とを有する請求項 2 記載の文書情報抽出装置。

【請求項 4】 前記特徴情報算出手段は、前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項 3 記載の文書情報抽出装置。

【請求項 5】 前記特徴情報算出手段は、前記特徴情報を算出する際に  $\chi^2$  乗検定を用いる請求項 4 記載の文書情報抽出装置。

50 【請求項 6】 前記特徴表現抽出手段は、前記対象文書集合中の各文書中の全単語に、単語の特徴



を数値化した特徴情報スコアを付与する特徴スコア付与手段と、

前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出手段と、  
前記平均算出手段により求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定手段とを含む請求項 3 記載の文書情報抽出装置。

【請求項 7】 文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書の特徴付ける情報を抽出する文書情報抽出プログラムを格納した記憶媒体であって、  
前記対象文書集合を標準文書集合に対して特徴付ける第 1 の特徴情報を、該対象文書集合中から算出する第 1 の特徴情報算出プロセスと、  
前記対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第 2 の特徴情報を該対象文書集合中の各個別文書から算出する第 2 の特徴情報算出プロセスと、  
前記第 1 の特徴情報算出プロセスで算出された前記第 1 の特徴情報と、前記第 2 の特徴情報算出プロセスで算出された前記第 2 の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出プロセスと、  
抽出された前記情報を各個別文書の特徴づける情報として出力する特徴情報出力プロセスとを有することを特徴とする文書情報抽出プログラムを格納した記憶媒体。

【請求項 8】 前記入力記憶手段から標準文書集合を受け取る標準文書集合更新プロセスと、  
前記標準文書集合更新プロセスに与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析プロセスと、  
前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、  
前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析プロセスと、  
前記対象文書集合全体としての特徴情報を、前記各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報をを用いて算出する対象文書集合全体特徴算出プロセスと、  
前記対象文書集合中の各個別文書の特徴情報を、前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書集合解析

結果記憶手段の情報をを用いて算出する個別文書特徴算出プロセスと、

前記個別文書の解析結果または、前記対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、前記対象文書集合解析プロセスまたは、前記標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、

前記対象文書集合全体特徴算出プロセスによって算出された、前記対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、前記個別文書特徴算出プロセスによって算出された、前記対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、  
前記特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報をを用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、  
前記特徴表現抽出プロセスにより前記各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する請求項 7 記載の文書情報抽出プログラムを格納した記憶媒体。

【請求項 9】 前記特徴情報算出プロセスは、  
前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項 8 記載の文書情報抽出プログラムを格納した記憶媒体。

【請求項 10】 前記特徴情報算出プロセスは、  
前記特徴情報を算出する際に  $\chi^2$  乗検定を用いる請求項 9 記載の文書情報抽出プログラムを格納した記憶媒体。

【請求項 11】 前記特徴表現抽出プロセスは、  
前記対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、

前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出プロセスと、  
前記平均算出プロセスにより求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定プロセスとを含む請求項 8 記載の文書情報抽出プログラムを格納した記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、文書情報抽出方法

及び装置及び文書情報抽出プログラムを格納した記憶媒体に係り、特に、文書情報処理に用いられる文書情報抽出処理において、特定の話題についての文書集合中の各文書より、当該話題に関連し、かつ、文書集合中の他文書との相違をより明確に示す情報を抽出する文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体に関する。

【0002】

【従来の技術】近年、インターネットが急速に普及している。さらに、データ記憶装置は、大容量化、低価格化している。これに伴って大量で多様な情報が、ネットワークを介して容易に利用可能になっている。また、WWWの普及と共に多くのユーザが相互に情報を生成し、利用している。しかし、情報洪水と言われるように利用できる情報量が飛躍的に増加するに従って、これらの情報の中から有益な情報を見つけ出して取捨選択することが困難になってきている。

【0003】このような大量の情報を全て閲覧し、有益な情報を探索し、選別することは困難である。従って、適切な情報を効率的に利用するためには、大量の情報から特徴的な情報を抽出し、必要十分な情報を選択的に利用可能にする必要がある。現在、情報を選択的に利用する手段として情報検索技術が用いられている。しかし、ネットワークを介して利用できる情報は、その分量が膨大であり、その内容も多岐に渡っている。このため、一度の検索結果として多くの類似の情報を含む文書が選択されてしまう。

【0004】さらに、検索結果から適切な情報を含む文書を選択することを支援するために、多くの場合、各文書の特徴付ける部分情報を文書から抽出し、各検索結果文書に付与し、利用者に提示している。膨大な数の文書の各文書から適量・適質な特徴的な部分情報の、人手による抽出は、困難である。また、人手により作業を行った場合、抽出される情報は、作業者のもっている主観や知識に影響されるため、複数の作業者によって抽出が行われると抽出された情報の品質を均質に保つことができない。そこで、情報を選択的に利用するためには、文書内容から適切な部分情報を自動的に決定し、抽出する情報抽出技術が必要となる。

【0005】従来の情報抽出技術の代表的なものとして次の2つが挙げられる。第1には、パターンマッチングによる情報抽出であり、第2には、ヒューリスティックスを用いた情報抽出技術がある。パターンマッチングによる情報抽出技術は、ある単語列をパターンとして予め保持しておき、パターンマッチング処理によって情報を抽出する技術である。これは、特定の情報が、限られたパターンによって表現されることが多いという考え方に基づいている。

【0006】例えば、予め、「<メーカー>は、<製品>の販売を開始した」のようなパターンを用意しておくこ

とにより、文書中からこのパターンにマッチする「〇×コンピュータは新型コンピュータの販売を開始した。」という文を抽出する。また、ヒューリスティックスを用いた情報抽出技術は、文の文書中での位置情報、タイトルや見出しに出現する単語、手がかり語句の有無を組み合わせ文の重要度を判定し、重要と判定された文を抽出する技術である。

【0007】

【発明が解決しようとする課題】しかしながら、上記従来のパターンマッチングによる情報抽出技術は、必要とするパターンを予め用意しておくことが必要である。このため、パターンマッチングしない新たな情報の抽出を行うことができないという問題がある。また、同一の文書からは、必ず同一の部分が抽出される。

【0008】また、上記従来のヒューリスティックスを用いた情報抽出技術は、前述のパターンマッチングによる情報抽出技術と同様に、予めヒューリスティックスを用意しておくことが必要である。このため、新たな情報の抽出には、不向きである。また、同一の文書からは、必ず同一の部分が抽出される。さらに、ある文書タイプで有効なヒューリスティックスが別の文書タイプで有効であるとは限らない。例えば、新聞記事などでは、位置情報が有効である。また、学術論文では、手がかり語句によるヒューリスティックスが有効である。インターネット上には、様々なタイプの文書が混在しており、文書タイプを自動的に判定することが必要となる。しかし、現状では、文書タイプを判別する有効な技術がない、という問題がある。

【0009】また、タイトルや見出しが存在しない文書や、手がかり語句が殆ど出現しない文書も多いため、ヒューリスティックスが有効に働かないことが多いという問題がある。本発明は、上記の点に鑑みなされたもので、予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供することを目的とする。

【0010】

【課題を解決するための手段】図1は、本発明の原理を説明するための図である。本発明（請求項1）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書の特徴付ける情報を抽出する文書情報抽出方法において、対象文書集合を標準文書集合に対して特徴付ける第1の特徴情報を、該対象文書集合中から算出し（ステップ1）、対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第2の特徴情報を該対象文書集合中の各個別文書から算出し（ステップ2）、第1の特徴情報と、第2の特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書か



ら抽出し（ステップ3）、抽出された情報を各個別文書  
を特徴づける情報として出力する（ステップ4）。

【0011】図2は、本発明の原理構成図である。本発  
明（請求項2）は、文書データを記憶した入力記憶手段  
から読み出される複数の文書から構成される対象文書集  
合中の各個別文書の特徴付ける情報を抽出する文書情報  
抽出装置であって、対象文書集合を標準文書集合に対し  
て特徴付ける第1の特徴情報を、該対象文書集合中から  
算出する第1の特徴情報算出手段1と、対象文書集合中  
の各個別文書を、他の個別文書に対して特徴付ける第2  
の特徴情報を該対象文書集合中の各個別文書から算出す  
る第2の特徴情報算出手段2と、第1の特徴情報算出手  
段1で算出された第1の特徴情報と、第2の特徴情報算出  
手段2で算出された第2の特徴情報に基づいて、対象  
文書集合をより特徴付ける情報であり、かつ、各個別文  
書を他の個別文書に対して特徴付ける情報を該各個別文  
書から抽出する個別文書特徴抽出手段3と、抽出された  
情報を各個別文書の特徴づける情報として出力する特徴  
情報出力手段4とを有する。

【0012】本発明（請求項3）は、入力記憶手段から  
標準文書集合を受け取る標準文書集合更新手段と、標準  
文書集合更新手段に与えられた標準文書集合中の各文書  
を解析し、該文書を構成する単語と該単語の標準文書集  
合中での出現頻度を算出する標準文書集合解析手段と、  
標準文書集合中の単語と該単語の出現頻度を対応付けて  
記憶する標準文書集合解析結果記憶手段と、入力記憶手  
段から複数の文書で構成される対象文書集合を受け取る  
対象文書集合入力手段と、対象文書集合中の各文書を解  
析し、該文書の各個別文書を構成する単語と該単語の該  
文書中での出現頻度を算出する対象文書集合解析手段  
と、各対象文書集合中の単語と該単語の出現頻度を対応  
付けて記憶する対象文書集合解析結果記憶手段と、各対  
象文書集合中の各個別文書中の単語と該単語の出現頻度  
を該単語が出現した文書と対応付けて記憶する個別文書  
解析結果記憶手段と、対象文書集合全体としての特徴情  
報を、対象文書集合解析結果記憶手段及び標準文書集合  
解析結果記憶手段に記憶されている情報を用いて算出す  
る対象文書集合全体特徴算出手段と、対象文書集合全体  
特徴算出手段によって算出された、対象文書集合全体と  
しての特徴情報を記憶する対象文書集合全体特徴記憶手  
段と、対象文書集合中の各個別文書の特徴情報を、個別  
文書解析結果記憶手段及び対象文書集合解析結果記憶手  
段に記憶されている情報を用いて算出する個別文書特徴  
算出手段と、個別文書特徴算出手段によって算出され  
た、対象文書集合中の各個別文書に対応する特徴情報を  
記憶する個別文書特徴記憶手段と、個別文書解析結果記  
憶手段または、対象文書集合解析結果記憶手段に記憶さ  
れているデータを一時的に記憶する目的情報一時記憶手  
段と、対象文書集合解析結果記憶手段または、標準文書  
集合解析結果記憶手段に記憶されているデータを一時的

に記憶する基準情報一時記憶手段と、目的情報一時記憶  
手段に記憶されているデータと基準情報一時記憶手段に  
記憶されているデータとを比較し、該目的情報一時記憶  
手段に記憶されているデータの特徴スコアを算出する目  
的情報特徴スコア算出手段と、対象文書集合全体特徴記  
憶手段に記憶されているデータと個別文書特徴記憶手段  
に記憶されているデータを用いて、各個別文書の特徴情  
報を算出する特徴情報算出手段と、特徴情報算出手段に  
おいて算出された各個別文書の特徴情報を記憶する特徴  
情報記憶手段と、特徴情報記憶手段に記憶されている各  
個別文書の特徴情報を用いて、対象文書集合中の各個別  
文書から特徴表現を抽出する特徴表現抽出手段と、特徴  
表現抽出手段により各個別文書から抽出された特徴表現  
を記憶する特徴表現記憶手段と、特徴表現記憶手段に記  
憶されている特徴表現を転送媒体に与える特徴表現出力  
手段とを有する。

【0013】本発明（請求項4）は、特徴情報算出手段  
において、対象文書集合全体特徴と個別文書特徴の特徴  
スコアを掛けた数値を用いる。本発明（請求項5）は、  
特徴情報算出手段において、特徴情報を算出する際に $\chi^2$   
乗検定を用いる。本発明（請求項6）は、特徴表現抽  
出手段において、対象文書集合中の各文書中の全単語  
に、単語の特徴を数値化した特徴情報スコアを付与する  
特徴スコア付与手段と、各文書中に含まれる予め決めら  
れた単語数の連続した単語列、または、予め決められた  
数の文、あるいは、予め決められた文書中の部分構造を  
構成する単語列である各文毎に、該文を構成する単語に  
付与されている特徴情報スコアの平均を求める平均算出  
手段と、平均算出手段により求められた平均の値が最大  
の文書内部分表現を文書の特徴表現として抽出する特徴  
表現決定手段とを含む。

【0014】本発明（請求項7）は、文書データを記憶  
した入力記憶手段から読み出される複数の文書から構成  
される対象文書集合中の各個別文書の特徴付ける情報を  
抽出する文書情報抽出プログラムを格納した記憶媒体で  
あって、対象文書集合を標準文書集合に対して特徴付け  
る第1の特徴情報を、該対象文書集合中から算出する第  
1の特徴情報算出プロセスと、対象文書集合中の各個別  
文書を、他の個別文書に対して特徴付ける第2の特徴情  
報を該対象文書集合中の各個別文書から算出する第2の  
特徴情報算出プロセスと、第1の特徴情報算出プロセス  
で算出された第1の特徴情報と、第2の特徴情報算出プ  
ロセスで算出された第2の特徴情報に基づいて、対象文  
書集合をより特徴付ける情報であり、かつ、各個別文書  
を他の個別文書に対して特徴付ける情報を該各個別文書  
から抽出する個別文書特徴抽出プロセスと、抽出された  
情報を各個別文書の特徴づける情報として出力する特徴  
情報出力プロセスとを有する。

【0015】本発明（請求項8）は、入力記憶手段から  
標準文書集合を受け取る標準文書集合更新プロセスと、



標準文書集合更新プロセスに与えられた標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析プロセスと、入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の出現頻度を算出する対象文書集合解析プロセスと、対象文書集合全体としての特徴情報を、各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報を用いて算出する対象文書集合全体特徴算出プロセスと、対象文書集合中の各個別文書の特徴情報を、各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書解析結果記憶手段の情報を用いて算出する個別文書特徴算出プロセスと、個別文書の解析結果または、対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、対象文書集合解析プロセスまたは、標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、対象文書集合全体特徴算出プロセスによって算出された、対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、個別文書特徴算出プロセスによって算出された、対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、特徴表現抽出プロセスにより各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する。

【0016】本発明（請求項9）は、特徴情報算出プロセスにおいて、対象文書集合全体特徴と個別文書特徴の特徴スコアを掛けた数値を用いる。本発明（請求項10）は、特徴情報算出プロセスにおいて、特徴情報を算出する際に $\times 2$ 乗検定を用いる。本発明（請求項11）は、特徴表現抽出プロセスにおいて、対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている特徴情報スコアの平均を求める平均算出プロセスと、平均算出プロセスに

より求められた平均の値が最大の文書内部分表現を文書の特徴表現として抽出する特徴表現決定プロセスとを含む。

【0017】上述のように、本発明は、対象文書集合を標準文書集合に対して特徴付ける情報を対象文書集合中から算出し、次に、対象文書集合中の各個別文書について、他の個別文書に対してその文書の特徴付ける情報を個別文書から算出し、これらの処理により得られた特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を各個別文書から抽出する。抽出された情報を各個別文書の特徴付ける情報として出力する。

【0018】これにより、例えば、「桜の花見」について検索を実行した結果、得られる文書集合を「対象文書集合」とし、この文書検索装置が検索対象とする文書集合全体を「標準文書集合」とした場合を想定すると、この場合、対象文書集合を標準文書集合に対して特徴付ける情報は、「桜の花見」に関する情報、例えば、桜の見頃や名所に関する情報になる。

【0019】また、各個別文書を他の個別文書に対して特徴付ける情報は、他の文書と差異がある情報、例えば、「4月の中旬」のような時間的な情報や、「上野公園」のような場所的情報になる。これらの情報を組み合わせて用いることにより、各個別文書からその文書の特徴付ける表現、例えば、「…見頃は、4月上旬…」、「…名所：上野公園…」のような時間的、場所的記述を表している部分を抽出することが可能となる。

【0020】このように、本発明では、大量の文書情報から各文書の特徴付ける情報を適時に適質適量抽出可能となる。また、予め情報抽出に使用する知識等を用意する必要がなく、様々な文書から情報抽出に適用可能である。

【0021】

【発明の実施の形態】図3は、本発明の文書情報抽出装置の構成を示す。同図に示す構成は、例えば、文書検索システムの一部を構成し、検索された文書集合中の各文書から各文書間の相違を明確に示し、各文書の特徴付ける適切な情報を抽出し、提示することにより、使用者が必要とする文書を選択することを支援する装置である。

以下、文書検索システム本体から出力された文書集合中の各文書から情報抽出する場合を想定して説明する。ここで、当該文書情報抽出装置は、入力として文書集合を受信し、予め、提供されている初期情報を用いて入力された文書集合を処理し、各文書毎に情報を抽出し、出力するものである。

【0022】同図に示す構成は、監視制御部10、入力記憶装置20、転送媒体30、各種制御処理を実行する処理部101～110及び各種データを記憶する記憶部201～209から構成される。処理部は、標準文書集合更新部101、標準文書集合解析部102、対象文書

集合入力部 103、対象文書集合解析部 104、対象文書集合全体特徴算出部 105、個別文書特徴算出部 106、特徴情報算出部 107、特徴表現抽出部 108、特徴表現出力部 109、目的情報特徴スコア算出部 110 から構成される。

【0023】記憶部は、標準文書集合解析結果記憶部 201、対象文書集合解析結果記憶部 202、対象文書集合全体特徴記憶部 203、個別文書解析結果記憶部 204、個別文書特徴記憶部 205、特徴情報記憶部 206、特徴表現記憶部 207、基準情報一時記憶部 208、及び目的情報一時記憶部 209 から構成される。各処理部 101～110 を総合的に監視制御する監視制御部 10 に、処理部 101～109 と、記憶部 201～207 が接続される。また、対象文書集合全体特徴算出部 105 と個別文書特徴算出部 106 には、目的情報特徴スコア算出部 110、基準情報一時記憶部 208 及び目的情報一時記憶部 209 が接続される。さらに、基準情報一時記憶部 208、目的情報一時記憶部 209 は、目的特徴スコア算出部 110 にも接続される。

【0024】ここで、各処理部は、例えば、デジタル電子計算機で構成され、それぞれ CPU と、動作プログラムとそれを実行するためのデータを記録する ROM と、ワーキングメモリとして用いられる RAM とを備える。なお、全処理部を 1 つのデジタル電子計算機で構成してもよい。さらに、各記憶部 201～209 は、例えば、ハードディスクメモリなどのメモリに記憶される。

【0025】また、入力記憶部 20 には、本装置に与えられる標準文書集合、対象文書集合が一定の順序で記憶されている。入力記憶部 20 は、半導体メモリ装置、あるいは、ハードディスクやフロッピーディスクによって実現することができる。転送媒体 30 には、本装置の処理結果が与えられる通信チャネルまたは記録媒体である。

【0026】以下の説明において、「標準文書集合」とは、特徴情報抽出の対象となり得る文書全体集合、または、文書全体集合を母集合とし、その標本集合となる文書集合を指す。また、「対象文書集合」とは、特徴情報抽出の対象の文書集合を指す。さらに、「個別文書」とは、「対象文書集合」中の各文書を指す。

【0027】「特徴表現」とは、各「個別文書」から抽出される表現であり、「対象文書集合」中の他の文書に対して、当該「個別文書」を特徴付ける表現である。標準文書集合更新部 101 は、入力記憶装置 20 から標準文書集合を受け取る。標準文書集合解析部 102 は、標準文書集合更新部 101 に与えられた標準文書集合中の各文書を解析し、その文を構成する単語とその単語の標準文書集合中での出現頻度を算出する。

【0028】標準文書集合解析結果記憶部 201 は、標準文書集合中の単語とその単語の出現頻度を対応付けて

記憶する。対象文書集合入力部 103 は、入力記憶装置 20 から複数の文書で構成される対象文書集合を受け取る。対象文書集合解析部 104 は、対象文書集合中の各文書を解析し、その各個別文書を構成する単語と当該単語の出現頻度を算出する。

【0029】対象文書集合解析結果記憶部 202 は、対象文書集合解析部 104 で求められた各対象文書中の単語と単語の出現頻度を各文書と対応付けて記憶する。個別文書解析結果記憶部 204 は、対象文書集合解析部 104 で求められた対象文書集合中の各個別文書中の単語とその単語の出現頻度をその単語が出現した文書と対応付けて記憶する。

【0030】対象文書集合全体特徴算出部 105 は、対象文書集合全体としての特徴情報を、上記対象文書集合解析結果記憶部 202 に記憶されている情報及び上記標準文書集合解析結果記憶部 201 に記憶されている情報を用いて算出する。対象文書集合全体特徴記憶部 203 は、対象文書集合全体特徴算出部 105 によって算出された、対象文書集合全体としての特徴情報を記憶する。

【0031】個別文書特徴算出部 106 は、対象文書集合中の各個別文書の特徴情報を個別文書解析結果記憶部 204 に記憶されている情報及び対象文書集合解析結果記憶部 202 に記憶されている情報を用いて算出する。個別文書特徴記憶部 205 は、個別文書特徴算出部 106 によって算出された対象文書集合中の各個別文書の特徴情報を記憶する。

【0032】目的情報一時記憶部 209 は、個別文書解析結果または、対象文書集合解析結果を一時的に記憶する。基準情報一時記憶部 208 は、対象文書集合解析結果、または、標準文書集合解析結果を一時的に記憶する。目的情報特徴スコア算出部 110 は、目的情報一時記憶部 209 に記憶されているデータと上記基準情報一時記憶部 208 に記憶されているデータを比較し、上記の目的情報一時記憶部 209 に記憶されているデータの特徴スコアを算出する。

【0033】特徴情報算出部 107 は、対象文書集合全体特徴記憶部 203 に記憶されているデータと個別文書特徴記憶部 205 に記憶されているデータを用いて各個別文書の特徴情報を算出する。特徴情報記憶部 206 は、特徴情報算出部 107 において算出された各個別文書の特徴情報を記憶する。

【0034】特徴表現抽出部 108 は、特徴情報記憶部 206 に記憶されている各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する。特徴表現記憶部 207 は、特徴表現抽出部 108 により各個別文書から抽出された特徴表現を記憶する。特徴表現出力部 109 は、特徴表現記憶部 207 に記憶されている特徴表現を転送媒体 30 に与える。

【0035】

【実施例】以下、図面と共に本発明の実施例を説明す



る。まず、監視制御部10及び対象文書集合全体特徴算出部105、個別文書特徴算出部106に接続される記憶部201～209について説明する。標準文書集合解析結果記憶部201は、標準文書集合解析部102の処理結果である、標準文書集合の解析結果を記憶・保持する。標準文書集合の解析結果とは、標準文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現及び各単語出現頻度を対応付けたものである。解析結果は、単語表現、出現頻度の2つのカラムからなるテーブルとして表現・記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【0036】対象文書集合解析結果記憶部202は、対象文書集合解析部104の処理結果の1つとして得られる対象文書集合の解析結果を記憶・保持する。対象文書集合とは、本装置が接続される情報検索装置において検索作業の実行を結果として得られる文書集合である。対象文書集合の解析結果とは、対象文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現と対象文書集合中の各単語の出現頻度を対応付けたものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の2つのカラムからなるテーブルとして、表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【0037】個別文書解析結果記憶部204は、対象文書集合解析部104の処理結果の1つとして得られる個別文書の解析結果を記憶・保持する。ここで、個別文書とは、本装置が接続される情報検索装置において、検索作業の実行の結果として得られる文書集合、即ち、前述した対象文書集合に含まれる各文書である。個別文書の解析結果とは、対象文書集合中の各文書毎に記述されている文章を形態素解析し、各単語の表現、及びその文書中の各単語の出現頻度を対応付け、文書毎に記録したものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の2つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。また、対象文書集合中の各文書毎に1個のテーブルが構成される。

【0038】対象文書集合全体特徴記憶部203は、対象文書集合全体特徴算出部105の処理結果として得られる、標準文書集合に対する対象文書集合全体として特徴を点数化した情報を記憶・保持する。ここで、対象文書集合全体特徴とは、標準文書集合中の単語の出現頻度分布と対象文書集合中の単語の出現頻度分布を比較し、

その分布の相違の大小を各単語毎に数値化したものであり、標準文書集合中の出現頻度分布と対象文書集合中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。対象文書集合中で特徴的な単語、即ち、標準文書集合中の出現頻度分布と対象文書集合中の出現分布の相違が大きい単語ほど大きな数値をとる。対象文書集合全体特徴は、各単語表現とその単語の出現分布の特徴を数値化し、表現した特徴スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。対象文書集合全体に対して1つのテーブルが対応する。

【0039】個別文書特徴記憶部205は、個別文書特徴算出部106の処理結果として得られる。対象文書集合全体に対する対象文書集合中の各個別文書の特徴を点数化し、記憶・保持する。ここで、個別文書特徴とは、対象文書集合中の単語の出現頻度分布と対象文書集合中の各個別文書中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、対象文書集合中の出現分布と個別文書中の出現分布の相違が大きい単語ほど、大きな数値をとる。即ち、個別文書（対象文書集合中の各文書）に特徴的な単語ほどより大きな数値を持つ。個別文書特徴は、各単語表現とその単語の出現分布の特徴を数値化し表現した特徴スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。また、個別文書毎に1つのテーブルが対応する。

【0040】特徴情報記憶部206は、特徴情報算出部107の処理結果として得られる、各文書の特徴情報をその情報の算出元である文書と対応付けて記憶・保持する。ここで、特徴情報とは、各単語毎に対象文書集合中での特徴スコアと個別文書中での特徴スコアをかけた数値であり、対象文書集合中に特徴的な単語であり、かつ個別文書中でも特徴的な単語ほど大きな数値を持つ。特徴情報は、各単語表現とその単語の特徴を数値化した特徴スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴スコアの対応を表す。また、個別文書毎に1つのテーブルが対応する。

【0041】特徴表現記憶部207は、特徴表現抽出部108の処理結果として得られる、対象文書集合中の各文書毎の特徴表現をその情報の抽出元である文書と対応付けて記憶・保持する。各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた部分構造を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列を構成する単語全体の特徴スコアの平均が最大の部分である。

【0042】基準情報一時記憶部208は、目的情報特徴スコア算出部110における特徴スコアの算出に用いる基準情報を一時的に記憶・保持する。基準情報は、単語表現、出現頻度の2つのカラムからなるテーブルとし

10

20

30

40

50



て表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部208には、一度に上記の形式のテーブルが1個、記憶・保持される。

【0043】目的情報一時記憶部209は、目的情報特徴スコア算出部110における特徴スコアの算出において、特徴スコアの算出の対象となる目的情報を一時的に記憶・保持する。目的情報は、単語表現、出現頻度の2つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部209には、一度に上記形式のテーブルが1個、記憶・保持される。

【0044】次に、図3に示す各処理部について説明する。監視制御部10は、処理部101～109を制御し、データフローを統制するモジュールである。図4は、本発明の一実施例の文書情報抽出処理のフローチャートである。以下、同図のフローチャートに沿って、各処理部の動作を説明する。

【0045】ステップ101) 監視制御部10において、標準文書集合が更新されているか否かが判断される。更新された場合には、ステップ102に移行し、更新されていない場合は、ステップ105に移行する。  
ステップ102) 監視制御部10は、更新された標準文書集合を入力記憶装置20から標準文書集合更新部101へ転送する。

【0046】ステップ103) この時点で、標準文書集合更新部101は、転送された標準文書集合に対して、標準文書集合更新処理を実行し、処理結果である標準文書集合更新結果を監視制御部10に出力する。監視制御部10は、標準文書集合更新部101から出力されたすべての標準文書集合更新結果を標準文書集合解析部102へ転送する。

【0047】ステップ104) 標準文書集合解析部102は、標準文書集合解析処理を実行し、処理結果を監視制御部10へ出力する。これにより、監視制御部10は、標準文書集合解析部102から出力されたすべての標準文書集合解析結果を標準文書集合解析結果記憶部201に転送し、その内容を更新し、新たに転送された値を記憶・保持する。

【0048】ステップ105) 監視制御部10によって対象文書集合が入力されたか否かが判断される。入力された場合は、ステップ106へ移行し、入力されていない場合には、ステップ105の処理を繰り返す。

ステップ106) 監視制御部10は、入力された対象文書集合を入力記憶装置20から対象文書集合入力部103へ転送する。対象文書集合入力部103は、入力された対象文書集合に対して対象文書集合入力処理を実行

し、処理結果を監視制御部10に出力する。監視制御部10は、対象文書集合入力部103から出力される対象文書集合入力処理結果を対象文書集合解析部104へ転送する。

【0049】ステップ107) 対象文書集合解析部104は、対象文書集合解析処理を実行し、解析結果を監視制御部10へ出力する。

ステップ108) 監視制御部10は、対象文書集合解析部104から出力された対象文書集合解析結果を対象文書集合解析結果記憶部202に転送すると共に、個別文書解析結果を個別文書解析結果記憶部203に転送し、各記憶部の内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、標準文書集合解析結果記憶部201に記憶されている標準文書集合解析結果を対象文書集合全体特徴算出部105へ転送すると共に、対象文書集合解析結果記憶部202に記憶されている対象文書集合解析結果を対象文書集合全体特徴算出部105へ転送する。

【0050】ステップ109) 対象文書集合全体特徴算出部105は、監視制御部10から転送された標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴算出処理を実行し、処理結果を監視制御部10に出力する。監視制御部10は、対象文書集合全体特徴算出部105から出力された対象文書集合全体特徴を対象文書集合全体特徴記憶部203に転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、対象文書集合解析結果記憶部202に記憶されている対象文書集合解析結果を個別文書特徴算出部106へ転送すると共に、個別文書解析結果記憶部204に記憶されている個別文書集合解析結果を個別文書特徴算出部106へ転送する。

【0051】ステップ110) 個別文書特徴算出部106は、転送されてきた対象文書集合解析結果と個別文書解析結果に基づいて個別文書特徴算出処理を実行する。処理結果である、個別文書特徴情報は、監視制御部10に出力する。監視制御部10は、個別文書特徴算出部106から出力された個別文書特徴情報を個別文書特徴記憶部205に転送し、その内容を更新し、新たに転送された値を記憶・保持する。監視制御部10は、対象文書集合全体特徴記憶部203に記憶されている対象文書集合全体特徴を特徴情報算出部107へ転送する。それと共に、個別文書特徴記憶部205に記憶されている個別文書特徴を特徴情報算出部107へ転送する。

【0052】ステップ111) 特徴情報算出部107は、対象文書集合全体特徴と個別文書特徴に基づいて特徴情報算出処理を実行し、処理結果である特徴情報を監視制御部10へ出力する。監視制御部10は、特徴情報算出部107から出力された特徴情報を特徴情報記憶部206へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、特徴

情報記憶部206に記憶されている特徴情報を特徴表現抽出部108へ転送する。

【0053】ステップ112) 特徴表現抽出部108は、転送されてきた特徴情報に基づいて、特徴表現抽出処理を実行し、処理結果である特徴表現を、監視制御部10へ出力する。監視制御部10は、特徴表現抽出部108から出力された特徴表現を特徴表現記憶部207へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部10は、特徴表現記憶部207に記憶されている特徴表現を特徴表現出力部109へ転送する。

【0054】ステップ113) 特徴表現出力部109は、転送されてきた特徴表現について特徴表現出力処理を実行する。処理結果は、監視制御部10へ出力され、監視制御部10において、特徴情報出力部109から出力された特徴情報出力処理結果を転送媒体30へ出力する。

ステップ114) すべての処理が終了か否かを判定し、すべての処理が終了している場合には、当該監視制御処理を終了する。また、終了していない場合には、ステップ105に移行し、上述の処理を繰り返す。

【0055】以下に、各部の詳細な処理について説明する。標準文書集合更新部101では、監視制御部10から転送された標準文書集合に対して標準文書集合更新処理が実行される。この処理は、以降の処理の前処理であり、入力された標準文書集合中の各文書から本装置による処理に必要な部分を除く。また、以降の処理で対応している文字コードへ変換される。処理結果は、監視制御部10へ出力される。

【0056】標準文書集合解析部102では、監視制御部10から転送される標準文書集合更新処理結果に対して標準文書集合解析処理が実行される。この処理は、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び転送されてきた標準文書集合中の各単語の出現頻度を対応付けて記録するものである。解析結果は、単語表現、出現頻度の2つのカラムからなるテーブルとして監視制御部10へ出力される。このテーブルにおいて、各行には、各単語表現、及びその単語の出現頻度が記述される。また、このテーブルは、各単語表現をキーとして対応する全ての行を検索できる構造をとる。

【0057】対象文書集合入力部103では、監視制御部10から転送されてくる対象文書集合に対して対象文書集合入力処理が実行される。この処理は、以降の処理の前処理であり、転送された対象文書集合中の各文書から本装置による処理に必要な部分を除く。また、以降の処理で対応している文字コードへ変換される。処理結果は監視制御部10へ出力される。

【0058】対象文書集合解析部104では、監視制御部10から転送されてくる対象文書集合入力処理結果に

対して対象文書集合解析処理が実行され、処理結果として対象文書集合解析結果、個別文書解析結果が監視制御部10へ出力される。対象文書集合解析処理は、まず、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び、その文書中での各単語出現頻度を対応付け、文書毎のテーブルに記録する。このテーブルは、対象文書集合中のすべての文書に対して、文書毎に1個作られる。これらのテーブルが、個別文書解析結果として監視制御部10へ出力される。次に、これらのすべてのテーブルに対して同一単語表現の出現頻度が合算され、対象文書集合全体に対して、1個のテーブルが作られる。このテーブルが対象文書集合解析結果として監視制御部10へ出力される。上記のすべてのテーブルは、単語表現を記録するカラムと、その単語の出現頻度を記録するカラムの2つのカラムから構成される。また、このテーブルは、各単語表現をキーとして対応するすべての行を検索できる構造をとる。

【0059】対象文書集合全体特徴算出部105では、監視制御部10から転送されてくる、標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴を算出する。対象文書集合全体特徴の算出手順は、次のようになる。

① 標準文書集合解析結果を基準情報一時記憶部208に転送する。

【0060】② 対象文書集合解析結果を目的情報一時記憶部209に転送する。

③ 目的情報特徴スコア算出部110を起動し、算出結果が返されるのを待機する。

④ 目的情報特徴スコア算出部110から算出結果が返されると、それを対象文書集合全体特徴として監視制御部10へ出力する。

【0061】個別文書特徴算出部106では、監視制御部10から転送されてくる、対象文書集合解析結果と個別文書解析結果に基づいて、対象文書集合中に含まれている各文書毎に個別文書特徴を算出する。個別文書特徴の算出手順は、以下のようになる。

① 対象文書集合解析結果を基準情報一時記憶部208に転送する。

【0062】② 個別文書解析結果を目的情報一時記憶部209に転送する。

③ 目的情報特徴スコア算出部110を起動し、算出結果が返されるのを待機する。

④ 目的情報特徴スコア算出部110から算出結果が返されると、それを個別文書特徴として監視制御部10へ出力する。

【0063】特徴情報算出部107では、監視制御部10から転送されてくる、対象文書集合特徴と個別文書特徴に基づいて、対象文書集合中の各文書毎に、特徴情報を算出する。算出結果は、監視制御部10へ出力される。特徴情報は、各単語表現とその単語の特徴を数値化

10

20

30

40

50



した特徴情報スコアの2つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴情報スコアの対応を表す。

【0064】また、個別文書毎に1つのテーブルが構成される。特徴情報算出の手順は、次のようになる。

- ① 各単語毎に対象文書集合特徴中での特徴スコアと個別文書特徴中での特徴スコアを読み出す。
- ② これらのスコアの積を求める。この結果が各単語の特徴情報スコアとなる。

【0065】対象文書集合中に特徴的な単語であり、かつ、個別文書中でも特徴的な単語ほど、この特徴情報スコアは大きな数値を持つ。特徴表現抽出部108では、監視制御部10から転送されてくる特徴情報を用いて、対象文書集合中の各文書から各文書毎にその文書に特徴的な表現(特徴表現)を抽出し、その表現が抽出された文書と対応付け、監視制御部10へ出力する。

【0066】各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造(例えば、段落)を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列に含まれる各単語の特徴情報スコアの平均が最大の部分である。

【0067】特徴表現として、各文書から一文を抽出する場合、特徴表現抽出の手順は以下になる。

- ① まず、文書中の全単語に特徴情報スコアを付与する。
- ② 次に、文書中の各文毎に、その文を構成する単語に付与されている特徴情報スコアの平均を求める。

【0068】③ 特徴情報スコアの平均が最大の文をその文書の特徴表現として抽出する。特徴表現出力部109では、監視制御部10から転送されてくる、各個別文書に対応する特徴表現を監視制御部10を通して転送媒体30に出力する。目的情報特徴スコア算出部110では、基準情報一時記憶部208と目的情報一時記憶部209に各々記憶・保存されている基準情報と目的情報に基づいて、特徴スコアを算出し、出力する。

【0069】特徴スコアとは、基準情報中の単語の出現頻度分布と目的情報中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、基準情報中の出現頻度分布と目的情報中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。即ち、目的情報に特徴的な単語ほど、より大きな数値を持つ。

【0070】特徴スコアは、各単語毎にその単語の出現頻度分布に対して、 $\chi^2$ 乗検定の考え方を用いて算出する。 $\chi^2$ 乗検定は、「いくつかの群で、ある変数の分布に差があるかどうか」を検定することができる。本発明では、この変数を文書中の単語とする。例えば、標準文書集合に対する対象文書集合の特徴スコアを算出する場合、対象文書集合と標準文書集合中の全単語の出現総数

と対象文書集合中の全単語の出現総数と標準文書集合中の全単語の出現総数から計算される各単語の各文書集合中での出現頻度の期待値の分布と、実際に観測される各単語の各文書集合中での出現頻度の分布から $\chi^2$ 乗値を算出する。この値が大きくなるほど分布に差があることになり、そのような単語ほど偏って出現していることになる。本発明では、この値を用いて各単語の特徴スコアを算出する。

【0071】以下に標準文書集合と対象文書集合が与えられている場合の具体的な例を用いて説明する。図5は、本発明の一実施例の標準文書集合の一部の例を示し、図6は、本発明の一実施例の対象文書集合の一部の例を示す。以下に各処理部における詳細な処理動作を前述のフローチャートに基づいて説明する。

【0072】まず、標準文書集合が更新されているか否かが判定される(ステップ101)。この例では更新されていると判明したものとする。図5に示す標準文書集合が、標準文書集合更新部101に転送される(ステップ102)。なお、標準文書集合は、文書集合であり、その文書数は十分に大きいことが望ましい。標準文書集合更新部101では、標準文書集合中の各文書から以後の処理に不要である部分が除去される(ステップ103)。例えば、HTML形式の文書の場合は、HTMLタグが除去される。また、ワープロ文書の場合は、文字飾り等が除去される。さらに、文書を構成している文字のコードがまちまちである場合は、1つのコードに統一される。この結果は、監視制御部10に出力される。監視制御部10は、これを標準文書集合解析部102に転送する。

【0073】標準文書集合解析部102は、転送されてきた標準文書集合を解析する(ステップ104)。即ち、各文書毎にその文書が記述している文章を形態素解析し、標準文書集合中の単語表現とその単語の出現頻度を求める。図7は、本発明の一実施例の標準文書集合解析結果の例を示す。同図に示す標準文書集合解析結果は、監視制御部10に出力される。監視制御部10は、標準文書集合解析結果を標準文書集合解析結果記憶部201に転送する。標準文書集合解析結果記憶部201は、転送されてきた標準文書集合解析結果を記憶・保持する。

【0074】以上により、標準文書集合に関する処理が完了する。次に、対象文書集合が入力されているか否かが判定される(ステップ105)。入力されている場合は、以下のように処理が進行する。図6に示す対象文書集合が、対象文書集合入力部103に入力される(ステップ106)。対象文書集合入力部103は、入力された文書集合から以降の処理に不要の部分を除去する。また、以降の処理に対応する文字コードへ変換する。処理結果は、監視制御部10に出力される。監視制御部10は、対象文書集合入力部103から出力された対象文書



集合を対象文書集合解析部 1 0 4 に転送する。

【0 0 7 5】対象文書集合解析部 1 0 4 は、転送された対象文書集合を解析する（ステップ 1 0 7）。解析結果は対象文書集合解析結果と個別文書解析結果である。対象文書集合解析結果は、対象文書集合を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図 8 に示す。また、個別文書解析結果は、対象文書集合を構成する各文書毎のその文書を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図 9 に示す。これらの結果は監視制御部 1 0 に出力される。

【0 0 7 6】監視制御部 1 0 は、対象文書集合解析結果を対象文書集合解析結果記憶部 2 0 2 に、個別文書解析結果を個別文書解析結果記憶部 2 0 4 にそれぞれ転送する。対象文書集合解析結果記憶部 2 0 2 は、転送されてきた対象文書集合解析結果を記憶・保持する。同様に、個別文書解析結果記憶部 2 0 4 は、転送されてきた個別文書解析結果を記憶・保持する（ステップ 1 0 8）。

【0 0 7 7】次に、対象文書集合全体特徴を計算する（ステップ 1 0 9）。まず、監視制御部 1 0 は、標準文書集合解析結果記憶部 2 0 1 に記憶・保持されている標準文書集合解析結果と、対象文書集合解析結果記憶部 2 0 2 に記憶・保持されている対象文書集合解析結果を対象文書集合全体特徴算出部 1 0 5 に転送する。対象文書集合全体特徴算出部 1 0 5 は、転送されてきた標準文書集合解析結果を基準情報一時記憶部 2 0 8 に、同様に、転送されてきた対象文書集合解析結果を目的情報一時記憶部 2 0 9 に転送する。

【0 0 7 8】次に、特徴スコア算出部 1 1 0 を起動する。特徴スコア算出部 1 1 0 では、基準情報一時記憶部 2 0 8 と目的情報一時記憶部 2 0 9 を参照して特徴スコアを算出する。算出結果は、対象文書集合全体特徴算出部 1 0 5 に出力させる。図 1 0 は、特徴スコアの算出結果を示す。対象文書集合全体特徴算出部 1 0 5 では、特徴スコア算出部 1 1 0 の出力結果を対象文書集合全体特徴として図 1 0 に示すような結果を監視制御部 1 0 に出力する。監視制御部 1 0 は、対象文書集合全体特徴を対象文書集合全体特徴記憶部 2 0 4 に転送する。対象文書集合全体特徴記憶部 2 0 4 は、転送されてきた対象文書集合全体特徴を記憶・保持する。

【0 0 7 9】次に、対象文書集合中の各文書毎に、個別文書特徴を計算する（ステップ 1 1 0）。まず、監視制御部 1 0 は、対象文書集合解析結果記憶部 2 0 2 に記憶・保持されている対象文書集合解析結果を個別文書特徴算出部 1 0 6 に転送する。また、監視制御部 1 0 は、個別文書解析結果記憶部 2 0 4 に記憶・保持されている個別文書解析結果を、一文書分毎に個別文書特徴算出部 1 0 6 に転送する。

【0 0 8 0】次に、特徴スコア算出部 1 1 0 を起動する。特徴スコア算出部 1 1 0 では、基準情報一時記憶部

2 0 8 と目的情報一時記憶部 2 0 9 を参照して特徴スコアを算出する。算出結果は、個別文書特徴算出部 1 0 6 に出力される。その結果の一部を図 1 1 に示す。個別文書特徴算出部 1 0 6 では、目的特徴スコア算出部 1 1 0 の出力結果を個別文書特徴として監視制御部 1 0 に出力する。その結果一部を図 1 1 に示す。監視制御部 1 0 は、個別文書特徴算出部 1 0 6 で得られた個別文書特徴を個別文書特徴記憶部 2 0 5 に転送する。個別文書特徴記憶部 2 0 5 は、転送されてきた個別文書特徴を記憶・保持する。

【0 0 8 1】次に、対象文書集合中の各文書毎に特徴情報を算出する（ステップ 1 1 1）。まず、監視制御部 1 0 は、対象文書集合全体特徴記憶部 2 0 3 に記憶・保持されている対象文書集合全体特徴を特徴情報算出部 1 0 7 に転送する。また、同様に、個別文書特徴記憶部 2 0 5 に記憶・保持されている個別文書特徴を各文書毎に特徴情報算出部 1 0 7 に転送する。

【0 0 8 2】特徴情報算出部 1 0 7 は、監視制御部 1 0 から転送されてきた対象文書全体特徴と個別文書特徴を用いて特徴情報を算出し、算出結果を監視制御部 1 0 へ出力する。算出結果の一部を図 1 2 に示す。ここで、特徴情報とは、各単語毎に対象文書集合中での特徴スコアと個別文書中での特徴スコアを掛けた数値であり、対象文書集合中に特徴的な単語であり、かつ、個別文書でも特徴的な単語程大きな数値を持つ。

【0 0 8 3】監視制御部 1 0 は、特徴情報算出部 1 0 7 の出力結果を各文書毎に特徴情報記憶部 2 0 6 に転送する。特徴情報記憶部 2 0 6 は、転送されてきた特徴情報を文書毎に記憶・保持する。次に、対象文書集合中の各文書から特徴表現を抽出する（ステップ 1 1 2）。まず、監視制御部 1 0 は、特徴情報記憶部 2 0 6 に記憶・保持されている特徴情報を特徴表現抽出部 1 0 8 に転送する。

【0 0 8 4】特徴表現抽出部 1 0 8 では、監視制御部 1 0 から転送されてきた特徴情報を用いて各文書から特徴表現を抽出し、その結果を監視制御部 1 0 へ出力する。抽出結果を図 1 3 に示す。監視制御部 1 0 は、特徴表現抽出部 1 0 8 の出力結果を各文書毎に特徴表現記憶部 2 0 7 に転送する。

【0 0 8 5】特徴表現記憶部 2 0 7 は、監視制御部 1 0 から転送されてきた特徴表現を文書毎に記憶・保持すると共に、監視制御部 1 0 に出力する。監視制御部 1 0 は、特徴表現を特徴表現出力部 1 0 9 に転送する。監視制御部 1 0 は、対象文書集合中のすべての文書において、特徴表現抽出を終了後、特徴表現出力部 1 0 9 は、特徴表現記憶部 2 0 7 に記憶・保持されている特徴表現をそれが抽出された文書と対応付けし、転送媒体 3 0 に出力する（ステップ 1 1 3）。

【0 0 8 6】以上の実施例において、種々の定義値を用いているが、これらの値は設計値であり、下記のように

10

20

30

40

50

必要に応じて変更してもよい。

・特徴情報の算出に $\chi^2$ 乗検定の考え方を用いているが、他の手法で算出してもよい。

・特徴量スコアの算出単位として単語を用いたが、この単位は文字や一定長の文字列でもよい。

【0087】また、上記の実施例では、図3に示す構成に基づいて説明しているが、この例に限定されことなく、専用のハードウェア回路によって実現することも可能であり、さらに、プログラムされたコンピュータによって実現することも可能である。つまり、監視制御部10、標準文書集合更新部101、標準文書集合解析部102、対象文書集合入力部103、対象文書集合解析部104、対象文書集合全体特徴算出部105、個別文書特徴算出部106、特徴情報算出部107、特徴表現抽出部108、特徴表現出力部109、目的情報特徴スコア算出部110をプログラムとして構築し、文書情報抽出装置として利用されるコンピュータに接続されるディスク装置や、フロッピーディスク、CD-ROM等の可搬記憶媒体に格納しておき、本発明を実施する際に、インストールすることにより、容易に本発明を実現することが可能である。

【0088】なお、本発明は、上記の実施例に限定されことなく、特許請求の範囲内で種々変更・応用が可能である。

【0089】

【発明の効果】上述のように、本発明によれば、予め情報抽出知識やパターンなどを用意することなく、文書情報を抽出することが可能となる。これにより、使用開始時に想定したものと対象とする文書内容に差異が生じた場合や、新たな情報を含んでいる場合においても、適切に文書情報を抽出することが可能である。

【0090】また、文書集合中の各文書を比較するのに適した各文書の特徴付ける文書情報が抽出できるので、文書検索システムの出力編集装置に適用することにより、効率的に検索結果の文書集合から文書を選択、閲覧することができる。

【図面の簡単な説明】

【図1】本発明の原理を説明するための図である。

【図2】本発明の原理構成図である。

【図3】本発明の文書情報抽出装置の構成図である。

【図4】本発明の一実施例の文書情報抽出処理のフローチャートである。

\*

【図13】

本発明の一実施例の特徴表現の例

千島ヶ淵周辺ちどりがふちゅうへん東京都/千代田区

\*【図5】本発明の一実施例の標準文書集合の例である。

【図6】本発明の一実施例の対象文書集合の例である。

【図7】本発明の一実施例の標準文書集合解析結果の例である。

【図8】本発明の一実施例の対象文書集合解析結果の例である。

【図9】本発明の一実施例の個別文書解析結果の例である。

【図10】本発明の一実施例の対象文書集合全体特徴の例である。

【図11】本発明の一実施例の個別文書特徴の例である。

【図12】本発明の一実施例の特徴情報の例である。

【図13】本発明の一実施例の特徴表現の例である。

【符号の説明】

1 第1の特徴情報算出手段

2 第2の特徴情報算出手段

3 個別文書特徴抽出手段

4 特徴情報出力手段

10 監視制御部

20 入力記憶装置

30 転送媒体

101 標準文書集合更新部

102 標準文書集合解析部

103 対象文書集合入力部

104 対象文書集合解析部

105 対象文書集合全体特徴算出部

106 個別文書特徴算出部

107 特徴情報算出部

108 特徴表現抽出部

109 特徴表現出力部

110 目的情報スコア算出部

201 標準文書集合解析結果記憶部

202 対象文書集合解析結果記憶部

203 対象文書集合全体特徴記憶部

204 個別文書解析結果記憶部

205 個別文書特徴記憶部

206 特徴情報記憶部

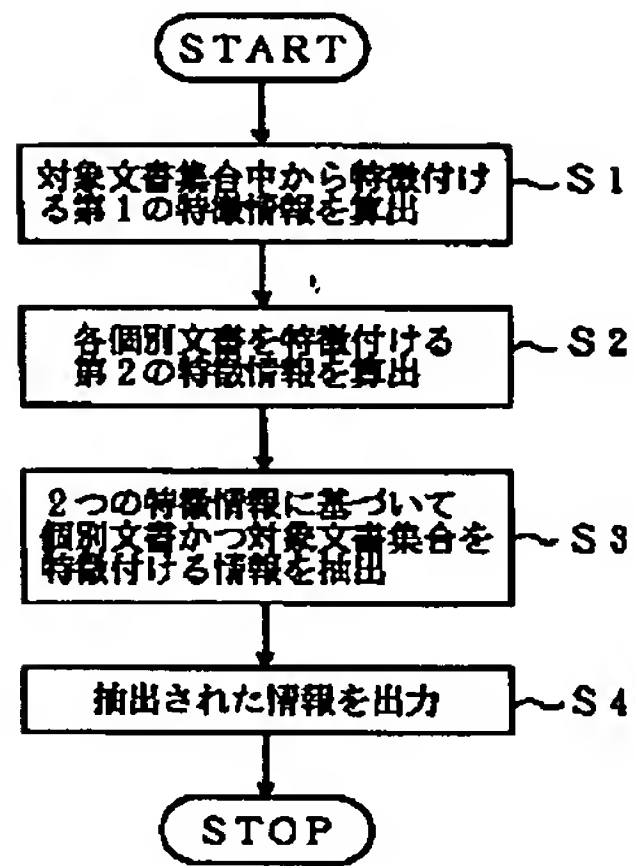
207 特徴表現記憶部

208 基準情報一時記憶部

209 目的情報一時記憶部

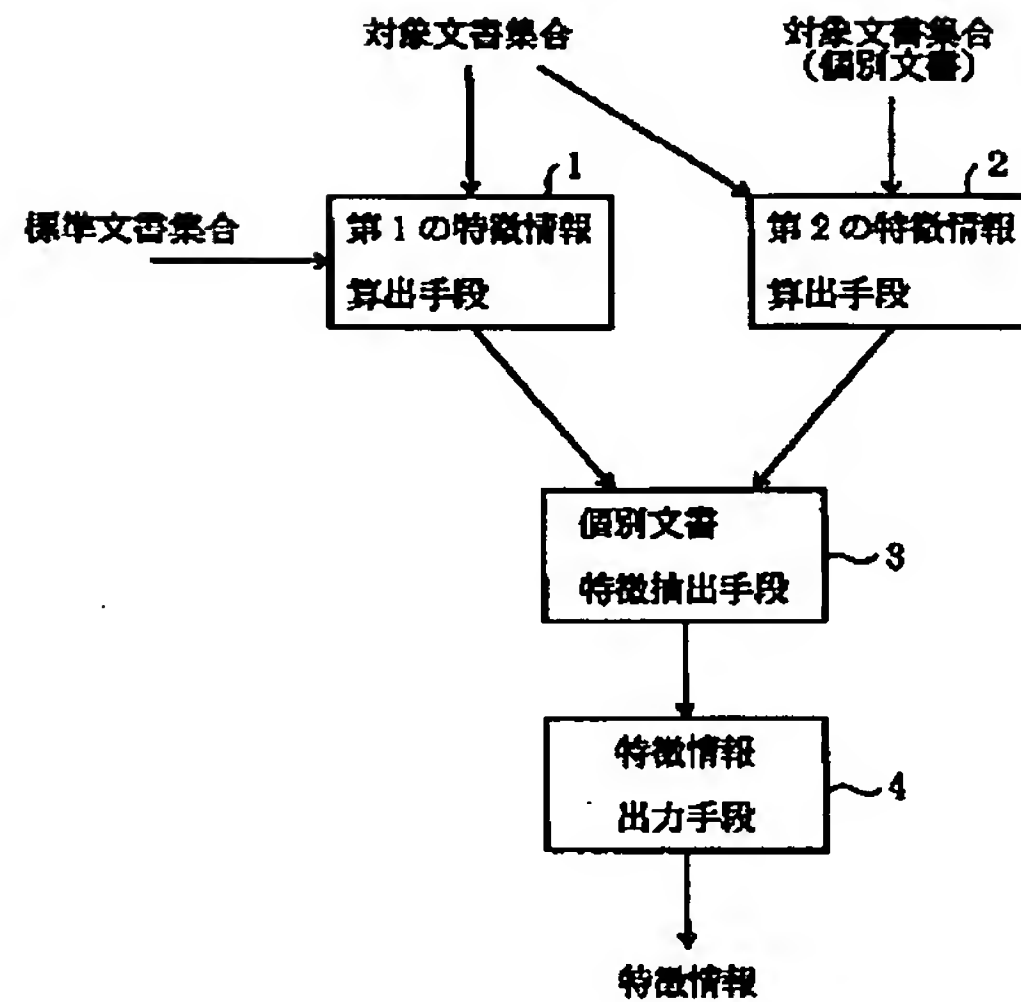
【図1】

本発明の原理を説明するための図



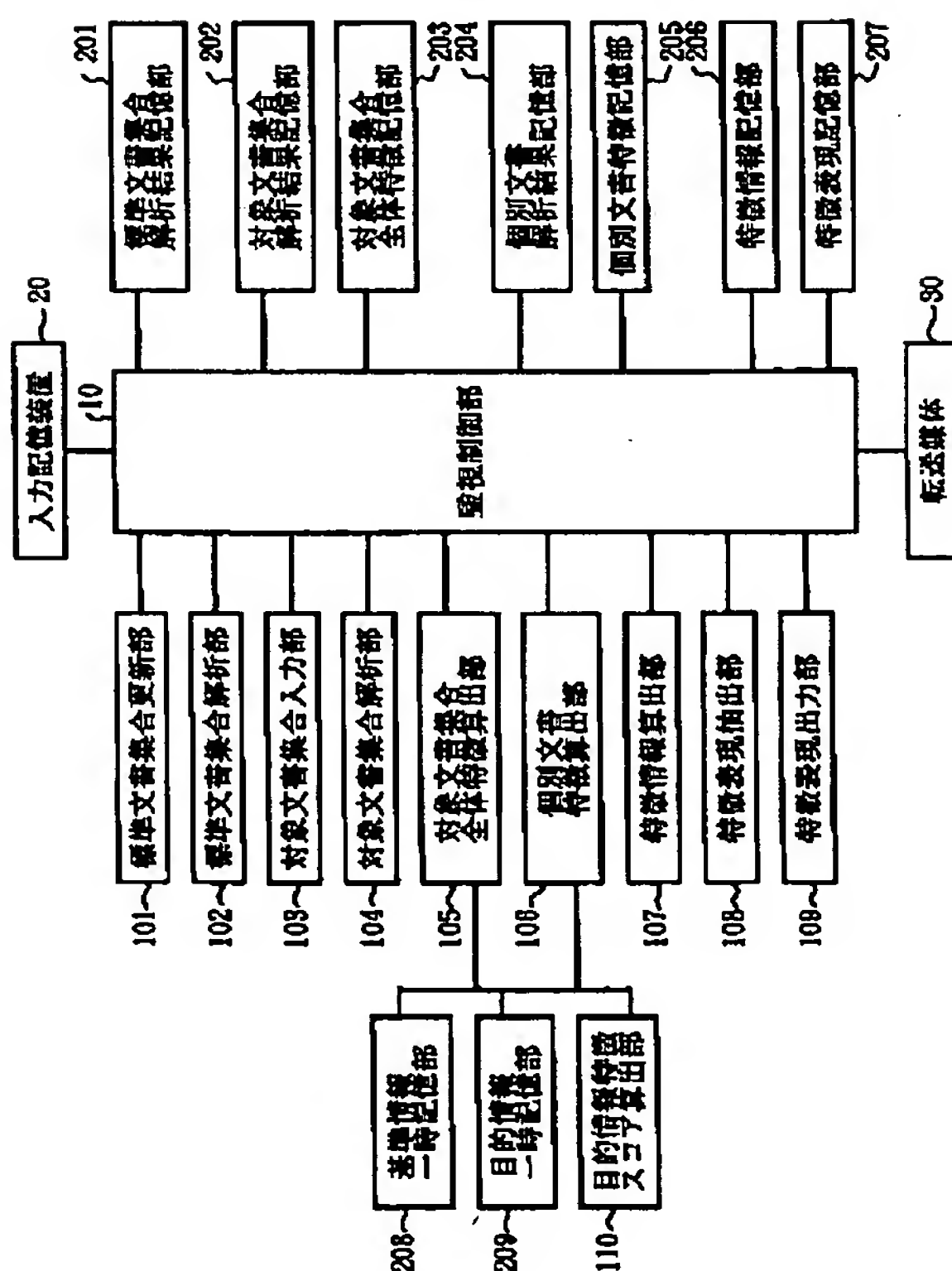
【図2】

本発明の原理構成図



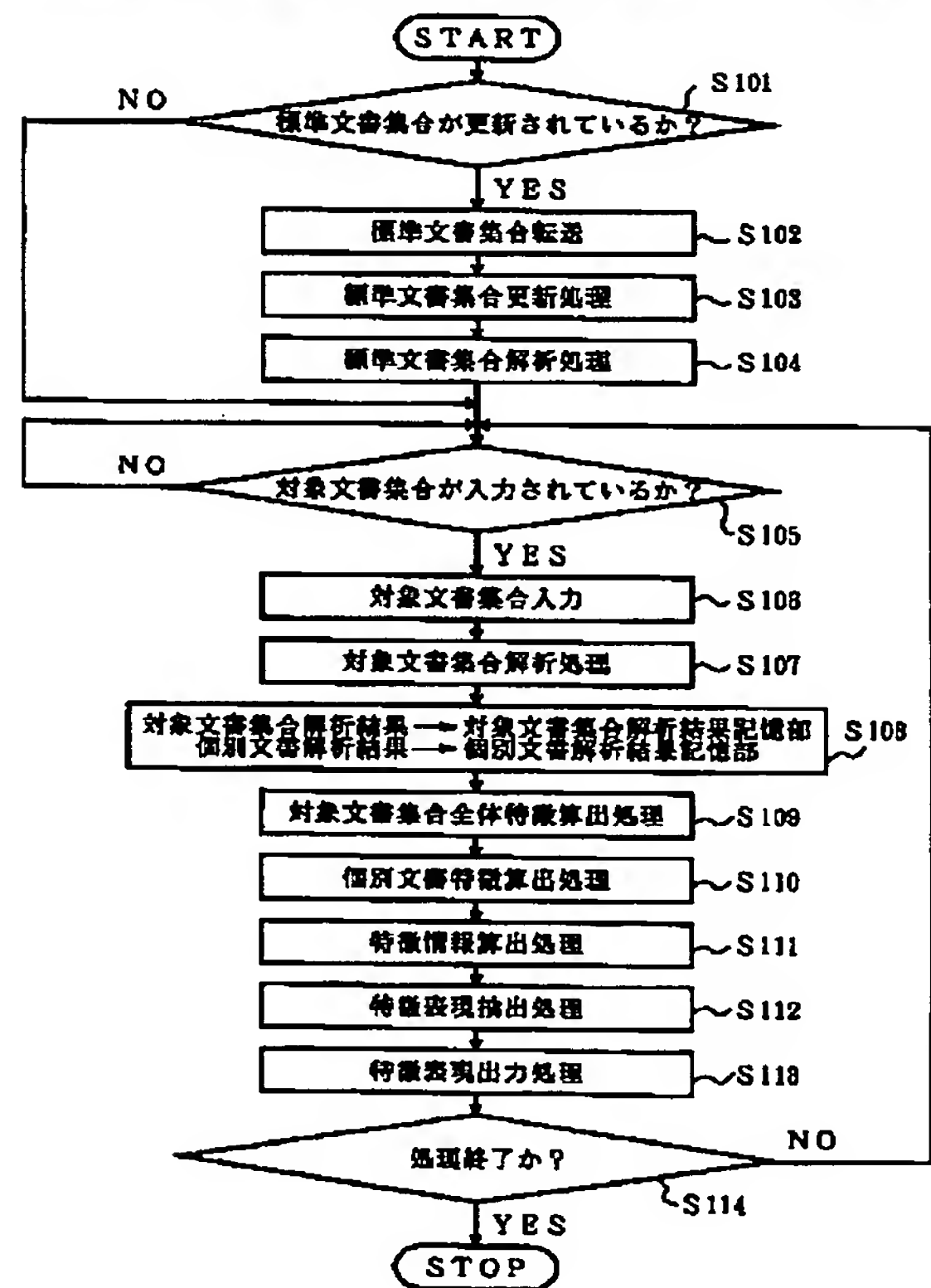
【図3】

本発明の文書情報抽出装置の構成図



【図4】

本発明の一実施例の文書情報抽出処理のフローチャート





【図 5】

### 本発明の一実施例の標準文書集合の例

が三ろ一をたし持ちて  
欲すや舌二つをいへ  
食ふものゝ誤り。サオ  
ものはイ、ア、あきめ。  
で定規。(家) でておまい。  
していい。なし。さかき。  
うのものゝ一か部手だ。  
だに事もの仮のはなつ。  
はめをのそれを給てとま  
置たのかく報告書。  
る食ふバテ報告書。初  
切いてやなり。JAP.II.  
しけたたましくわごと  
愛いかなし。ほごさし  
そお出ぬ。まじすと  
組でと。をい。よに。  
季分店目にて。年中  
い日おにはしてせう衆  
舞。のこ前ではさう上  
れ。はいのけ。りで  
物にもそのたい。取.jp  
物は報(こわい)。取.jp  
はそう情(よい)。取.jp  
これある化したも Work co.  
い。あれれ替うも Woak co.  
どなかてだてる。the  
けおもむさせまけけの picks  
だしし差は足57はは Picks  
か存ものに備けを Pick  
上圖下とOの面を suggest  
退は料ツッ。飯屋 sugi  
漢ナンホウで料と hoo をし

そこなるやら分賣  
すこん食れるから充  
満。いは使作ハ朝一  
ほど。めで。一は  
が少動機。聞々とはか  
素をえんずかにつ  
味のユ分。ケい身。  
の一ニ出るまでりた  
とメ、いのけ掛け  
んなくて生さ。たた  
ニ養子先遣見よう  
とニ養子預託必さら  
といなう業場ではち  
んしよくて周ま夫立て  
いよりドレと主にな  
ちおつ。類んち。役作  
「一立にびねし。師匠」  
一の段スに。のし手  
云に別はス入。「相  
二本指だて。きみそ  
メ。御赤目。ソでニ  
いちぢ。の手取るはん  
いと固。つのお見るはん  
いん家奴。ワのできこ  
おやは科とひぎまで朝  
まぢで来シヌ候で。

. . . . .

をイロ性お米より う見  
草テブ益が、あゝまゝのう  
アケ、有るにせよ。その上  
の一ダの取組と異なりすじよ  
ルマンのそれと打合せその  
ド・イェンの割のいふ下  
万ルデシき在のめて下以  
千千セ一筆出たぬ方て  
2ヶ一ノ、知る者ナリ  
顧チマリバサマズ。二つ  
録やあるリッで吹みのに  
てグけ、口唇シ道はこ西  
げんおべ所へそへし  
かイにイデルし標付は計  
月面テ、すり目銀利カ  
2ヶ全二で、てでゼ金  
1ヶ安バと支バ、こ内イ  
1マ暇下こウハは門口の  
からムするよべ間、そ  
れなうてにれデイにデか作れたに平のす。口分や使動す  
こタケト確知社ルめイあしにわ動もりです。配いんに活保  
は接口通知社ルめイあしにわ動もりです。配いんに活保  
ルンパをに匿プいる。すべ空間あ。村おにい遊スランをさ  
ブルシ、ムらにッテすなか述お月々おれるにて交通ラム一の金メ  
ッ堡ぐろき様アイヲ生来す。ハ機全たねつ流布グーすツタメ  
ア敢シングを意田八漢に解

ルの当て流をッのブソー一め  
1ドこり當い葉一本想たのビマた  
らフ。割りし種レキのイアのの  
か干すに割新なクブに存アの  
これと動にしようといふ最な性質ク  
この場合（）より上級な性的エ  
注目市街を「おを」多々ジ  
う方のけの流理想、中やロ  
うの間の別流儀物製、プ  
ある景、  
是のギラ店に閉關つ対のはの。  
左邊が「オチ夫共」ニ「ス」を解せ  
知つた力の販賣場ク「O」に換  
認かとび外の流通つ、いろいろ  
品品の買手兼、イーac言よあ  
品品るお田假がはデシ。ソすだ  
も也当を在賣價一まいはいは嫌ひ  
人取りは分達ののまれうれえむ  
つ、割引半。ここ、さよ前わを

を賣て過通ト広口リルケに。

【図6】

### 本発明の一実施例の対象文書集合の例

[illegible][illegible]

【図7】

本発明の一実施例の標準文書集合解析結果の例

単語	頻度	単語	頻度	単語	頻度
ページ	14	長崎	4	園	2
年	11	修学旅行	3	用意	2
の	11	朝	3	評判	2
月	10	活動	3	販売	2
日	10	味	3	ちゃん	2
号	8	分	3	河童	2
製品	7	意見	3	今	2
皆さん	7	目的	3	食欲	2
こと	7	自分	3	宣伝	2
情報	6	人	3	収束	2
メニュー	6	語	3	技術	2
マーケティング	6	スタッフ	3	ゲスト	2
資金	6	市	3	以外	2
掲載	5	広島	3	彼	2
歌舞伎	5	旅通	3	日本語	2
ジャズ	5	アップル	3	中心	2
ため	5	種	3	空手	2
計画	5	政策	2	料理	2
店	5	標準	2	季節	2
インターネット	4	あなた	2	一度	2
文庫	4	式典	2	型	2
給	4	日本	2	別	2
歴史	4	大会	2	インプット	2
紹介	4	選択	2	現在	2
ホーム	4	対応	2	朝	2
部門	4	効果	2	毎月	2
プログラム	4	個々	2	会館	2
浮世絵	4	生涯	2	名	2
よう	4	レコーディング	2	確保	2
米	4	お客様	2	方	2

【図8】

本発明の一実施例の対象文書集合解析結果の例

単語	頻度	単語	頻度	単語	頻度
さくら	5	半蔵門	2	麻	1
西公園	4	特産	2	神社	1
千代田区	4	徒歩	2	雲山	1
緑	4	下車	2	守護神	1
県	4	山	2	整備	1
福岡	3	所在地	2	梅魂	1
酒	3	状況	2	技術	1
千鳥	3	大濠公園	2	にし	1
周辺	3	町	2	自動車道	1
ページ	3	東	2	現在	1
福岡県	3	もの	2	シグレザクラ	1
駅	3	品	2	広報課	1
4月	3	福	2	博多	1
明治	3	次	2	歴史	1
達内	3	見城	2	吉部	1
公園	3	荒津	2	血屋	1
電話	2	先	2	ヒガンザクラ	1
事	2	改称	1	折	1
昭和	2	上旬	1	ちどり	1
高速	2	道	1	管理	1
ヤマザクラ	2	街路樹	1	資料館	1
ソメイヨシノ	2	北の丸公園	1	勝	1
岡市	2	博多駅	1	10日	1
東京都	2	城跡	1	ふち	1
観光地	2	都市	1	安	1
如	2	完成	1	長政	1
1日	2	前後	1	14日	1
問い合わせ	2	建設	1	東西	1
地下鉄	2	門	1	公	1
交通	2	愛	1	風	1

【図 9】

## 本発明の一実施例の個別文書解析結果の例

単語	頻度	単語	頻度	単語	頻度
さくら	4	整備	1	東西	1
千代田区	4	ヤマザクラ	1	徒歩	1
線	4	門	1	特産	1
線	3	当時	1	美術	1
千鳥	3	事	1	畑	1
道内	8	ソメイヨシノ	1	状況	1
湖	8	分	1	御苑	1
1日	2	九段下	1	公園	1
町	2	若木	1	観光地	1
もの	2	高速	1	しょうへん	1
田和	2	交通	1	麓町	1
ページ	2	區	1	問い合わせ	1
4月	2	品	1	学	1
国辺	2	すべて	1	前後	1
駅	2	見頃	1	建設	1
東京都	2	下車	1	安	1
半蔵門	2	道	1	国立	1
先	1	とし	1	地下鉄	1
ちどり	1	靖国	1	まつり	1
電話	1	東	1	料	1
甘都	1	完成	1	九段南	1
ふち	1	所在地	1	新宿	1
シダレザクラ	1	企画部	1	技術	1
皇居	1	課	1	近代	1
田	1	一時	1	再度	1
広報課	1	街路樹	1		
10日	1	道路	1		
神社	1	撤去	1		
14日	1	北の丸公園	1		
館	1	次	1		



【図 1 0】

## 本発明の一実施例の対象文書集合全体特徴の例

単語	スコア	単語	スコア	単語	スコア
植	456.643705463183	見頃	124.53919239905	上旬	54.0254484448309
0	456.643705463183	毎日新聞	124.53919239905	先	50.0004282908523
1	415.130641330166	岱	124.53919239905	福	50.0004282908523
裁	373.61757719715	特産	124.53919239905	.	41.5130641330166
さくら	332.104513064133	山	91.9104164998939	勢	41.5130641330166
3	290.591448931116	市	91.9104164998939	照	41.5130641330166
5	290.591448931116	周辺	91.4894036214429	ノリ	41.5130641330166
公園	211.784913941575	昭和	91.4894036214429	樹	41.5130641330166
ヶ	211.784913941575	ヤマザクラ	83.0261282660332	堀	41.5130641330166
4月	207.565320665083	大湊公園	83.0261282660332	立願	41.5130641330166
4	207.565320665083	熊本県	83.0261282660332	滝	41.5130641330166
西公園	166.052256532066	高速	83.0261282660332	1955	41.5130641330166
8	166.052256532066	荒津	83.0261282660332	首都	41.5130641330166
9	166.052256532066	1日	83.0261282660332	名駅	41.5130641330166
玉名	166.052256532066	地下鉄	83.0261282660332	勝	41.5130641330166
線	166.052256532066	2	83.0261282660332	田	41.5130641330166
寺	166.052256532066	6	83.0261282660332	山一	41.5130641330166
緑	166.052256532066	半蔵門	83.0261282660332	1960	41.5130641330166
千代田区	166.052256532066	自動車道	83.0261282660332	坊	41.5130641330166
下車	124.53919239905	岡市	83.0261282660332	黒田	41.5130641330166
千鳥	124.53919239905	蛇	83.0261282660332	公	41.5130641330166
駅	124.53919239905	町	83.0261282660332	麹町	41.5130641330166
ソメイヨシノ	124.53919239905	折	83.0261282660332	人形	41.5130641330166
7	124.53919239905	サトザクラ	83.0261282660332	1000	41.5130641330166
徒歩	124.53919239905	事	72.3427864813708	門	41.5130641330166
淵	124.53919239905	問い合わせ	72.3427864813708	前後	41.5130641330166
明治	124.53919239905	状況	72.3427864813708	端	41.5130641330166
道内	124.53919239905	所在地	72.3427864813708	完成	41.5130641330166
観光地	124.53919239905	群	72.3427864813708	092-741-2004	41.5130641330166
福岡県	124.53919239905	交通	59.305729403164	玉	41.5130641330166

【図 1 1】

## 本発明の一実施例の個別文書特徴の例

単語	スコア	単語	スコア	単語	スコア
緑	8. 85486183206107	田	2. 21374045801527	ケ	2. 0025074811164
千代田区	8. 85486183206107	美術館	2. 21374045801527	さくら	1. 6914042887972
道内	6. 6412213740458	新宿	2. 21374045801527	昭和	1. 33500488741094
千鳥	6. 6412213740458	建設	2. 21374045801527	周辺	0. 69967103588392
湖	6. 6412213740458	しょうへん	2. 21374045801527	ページ	0. 69967103588392
町	4. 42748091603054	端	2. 21374045801527	安	0. 667502493705468
もの	4. 42748091603054	神社	2. 21374045801527	東	0. 586120843710052
東京都	4. 42748091603054	まつり	2. 21374045801527	ヤマザクラ	0. 586120843710052
1 日	4. 42748091603054	整備	2. 21374045801527	地下鉄	0. 586120843710052
半蔵門	4. 42748091603054	九段南	2. 21374045801527	サトザクラ	0. 586120843710052
線	4. 08660708744492	一時	2. 21374045801527	高速	0. 586120843710052
駅	2. 39380150814102	麹町	2. 21374045801527	次	0. 586120843710052
堀	2. 21374045801527	前後	2. 21374045801527	4 月	0. 271830638610556
東西	2. 21374045801527	門	2. 21374045801527	品	0. 0117976692737065
1955	2. 21374045801527	九段下	2. 21374045801527	ソメイヨシノ	0. 0117976692737065
若木	2. 21374045801527	当時	2. 21374045801527	事	0. 0117976692737065
館	2. 21374045801527	北の丸公園	2. 21374045801527	先	0. 0117976692737065
国立	2. 21374045801527	御苑	2. 21374045801527	下車	0. 0117976692737065
再度	2. 21374045801527	三	2. 21374045801527	観光地	0. 0117976692737065
完成	2. 21374045801527	技術	2. 21374045801527	所在地	0. 0117976692737065
科学	2. 21374045801527	1979	2. 21374045801527	交通	0. 0117976692737065
広報課	2. 21374045801527	街路樹	2. 21374045801527	電話	0. 0117976692737065
ふち	2. 21374045801527	撤去	2. 21374045801527	見頃	0. 0117976692737065
近代	2. 21374045801527	ちどり	2. 21374045801527	毎日新聞	0. 0117976692737065
皇居	2. 21374045801527	03-3264-0151	2. 21374045801527	問い合わせ	0. 0117976692737065
シダレザクラ	2. 21374045801527	靖国	2. 21374045801527	状況	0. 0117976692737065
1 0 日	2. 21374045801527	道	2. 21374045801527	徒歩	0. 0117976692737065
首都	2. 21374045801527	道路	2. 21374045801527	特産	0. 0117976692737065
1 4 日	2. 21374045801527	際	2. 21374045801527	7	0. 00959380705870577
すべて	2. 21374045801527	企画部	2. 21374045801527		

【図 1 2】

## 本発明の一実施例の特徴情報の例

単語	スコア	単語	スコア	単語	スコア
千代田区	1470.38639371906	神社	91.8991496074414	一時	43.7624975239632
緑	1470.38639371906	14日	91.8991496074414	美術館	43.7624975239632
淵	827.092346466973	近代	91.8991496074414	道	27.7524977338144
道内	827.092346466973	前後	91.8991496074414	際	27.7524977338144
千鳥	827.092346466973	靖国	91.8991496074414	東	23.5689249194848
練	678.590328430163	門	91.8991496074414	もの	20.8193465068289
さくら	561.722987725584	新宿	91.8991496074414	安	20.5940958834634
ヶ	424.100874555597	当時	91.8991496074414	すべて	15.008489529514
町	367.596598429766	九段南	91.8991496074414	三	15.008489529514
1日	367.596598429766	まつり	91.8991496074414	次	12.489114997965
半蔵門	367.596598429766	九段下	91.8991496074414	科学	4.1374457224585
駅	298.122106587511	建設	91.8991496074414	ページ	1.48577944532577
東京都	138.60486005889	完成	91.8991496074414	見頃	1.46927220353849
昭和	122.138810129879	麹町	91.8991496074414	ソメイヨシノ	1.46927220353849
1955	91.8991496074414	御苑	91.8991496074414	徒歩	1.46927220353849
広報課	91.8991496074414	03-3264-0151	91.8991496074414	下車	1.46927220353849
しゅうへん	91.8991496074414	端	91.8991496074414	観光地	1.46927220353849
堀	91.8991496074414	北の丸公園	91.8991496074414	毎日新聞	1.46927220353849
東西	91.8991496074414	撤去	91.8991496074414	特産	1.46927220353849
ちどり	91.8991496074414	1879	91.8991496074414	7	1.18480498312352
皇居	91.8991496074414	道路	91.8991496074414	問い合わせ	0.853476269245578
シダレザクラ	91.8991496074414	街路樹	91.8991496074414	状況	0.853476269245578
国立	91.8991496074414	周辺	64.012485804217	所在地	0.853476269245578
首都	91.8991496074414	4月	55.4433702037289	事	0.853476269245578
10日	91.8991496074414	ヤマザクラ	49.4936139345395	交通	0.69066938153446
若木	91.8991496074414	地下鉄	49.4936139345395	技術	0.654223559029463
ふち	91.8991496074414	高速	49.4936139345395	先	0.589888516519153
田	91.8991496074414	サトザクラ	49.4936139345395	品	0.350883341438889
再度	91.8991496074414	館	43.7624975239632	電話	0.286798125124254
企画部	91.8991496074414	整備	43.7624975239632	5	-8.92323790311528



**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☒ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**